Open Data Operations for COVID-19

October 7, 2020

Aaron Katz

Large Scale Analytics Systems Group
Asymmetric Operations Sector
Johns Hopkins University Applied Physics Laboratory
A Brief History
In the Beginning

It started with coffee

And ended with a Tweet heard round the world
The Dashboard was rapidly picked up by US and international press.
Early Data Operations

“Authoritative” Data Sources

Crowd Sourced Data

Manual Data Curation

Manual Dashboard & Data Push

Fully manual process
Not scalable – Limited Sharing
The Early Partnership
The PAI Landscape Feb 2020
Open Source Surveillance

• Survive
  - Help Lauren Garnder and Frank Dong survive the data surge:
    ▪ Help ID trusted data sources
    ▪ Help Triage recommendations and comments
    ▪ Identify opportunities for process automation and prototype solutions
  - Help improve information sharing with key partners
    ▪ Backend augmented to support
  - Other duties as assigned

• Thrive
  - Help get surveillance picture closer to real-time while preserving quality
  - Create a configurable infrastructure that allows us to rapidly stand-up capability in any crisis
The PAI Landscape Feb 2020
Disinformation Shapes the Battlespace

Julia Carrie Wong @juliacarriew · 7h
As coronavirus misinformation spreads, the press points the finger at social media and trolls. But professional media outlets have been a conduit of disinformation.

That racist “bat soup” canard? Thank British newspapers for spreading around.

CNN @CNN · Jan 28
There's been a run of surgical masks in the US because of the coronavirus scare. You don't need them, physicians say. cnn.it/2GzpKyM

Chris Cyrek @SquarsjawCity · Jan 27
#Wuhan BSL-4 Research Laboratory was using a bat-derived SARS-like coronavirus to test it's disinfectant procedures in 2018. Genetic sequencing of 2019 Novel CoV suggests the current virus is extremely similar. Proximity of the Research Lab to outbreak origin deeply troubling.

CaliCal2000 @CaliCal2000 · Jan 30
* 300 US military personnel arrived in Wuhan for the Military World Games on October 19. The first coronavirus case appeared two weeks later, on November 2. Coronavirus incubation period is 14 days.*

-China #bioweapon

Coronavirus: the Dark Side
Vioxx killed 500,000 Americans: a toll that could have been reduced by 90% had the FDA issued a ... unz.com
This has Real Consequences

New York’s Orthodox Jewish community is battling measles outbreaks. Vaccine deniers are to blame.

Rockland County and New York City have declared

Tiny Samoa has had nearly 5,000 measles cases. Here’s how it got so bad.

This island nation saw 70 measles deaths so far this year, most in children under age 5.

Russia trolls 'spreading vaccination misinformation' to create discord

Just before my phone gets taken 1000mg vit C sodium ascorbate taken dissolved in 1/4 cup water every 3 hours. This will save your kids. They should be on their feet in 16 hours. Then start vitamins A for repair of damaged tissue.

Save our kids. Nobody should have died. This is the greatest crime against our people by our own people.

God Bless Samoa. Please stop this madness.
The PAI Landscape Feb 2020
Championing the Truth Can Change Behaviors and Save Lives

How was it contained?
By educating West Africans on how Ebola is transmitted, and by aggressive efforts to isolate and treat the infected. When the outbreak began 16 months ago, health care workers in Liberia, Guinea, and Sierra Leone were initially fighting blind. Doctors and nurses had to learn from scratch how to treat Ebola — the odds of survival dramatically increase with early, effective symptom management — and how to contain it. Helped by more than 10,000 volunteers from around the world, health workers gradually taught people to avoid unnecessary physical contact, to go to a clinic the moment they displayed symptoms, and to forgo the traditional ritual of washing corpses — a practice that accelerated the spread of the disease.
"The best way to fight Ebola," says Joseph Boye Cooper, a volunteer worker in Liberia, "is to prevent it."
Back on Track
COVID Data Pipeline

Global Web Facing Data on Cases, Deaths, and Recoveries

Curated Public Data Products

GIS & Web Services

Optimized JHU Infrastructure

Public REST APIs

Public Consumers
COVID Data Pipeline - Initial Architecture

- Global Sitrep
- National Reporting
- State & Local Reporting

"Authoritative" Data Sources

Crowd Sourced Data

Expanded Data Sourcing

Collection → Curation → Production

Automated Product Generation

Open Data Sharing

Automated Updates

Automated Data Collection – Manual Error Recovery
Expanding Use
Led to Expanding Demands
And Massive Growth in Data Requests

4.5 Billion requests/day

Stable U.S. County Reporting
Why this is hard
(Hint: Standards Matter)

• Data reporting is inconsistent and constantly changing

• The scale of data is such that it requires methods for collection and collation that are both nuanced and complex

• Definitions and collection capabilities are inconsistent across counties, states, and countries

Simultaneously addressing both source instability globally and operating near a zero error rate requires extensive engineering and effort
Refined Gen 1 Architecture

- Global Sitreps
- National Reporting
- State & Local Reporting
- "Authoritative" Data Sources
- Crowd Sourced Data
- Collection
- Curation
- Production

Expanded Data Sourcing

Automated Product Generation

Open Data Sharing

Automated Updates

Parallel Testing Pipeline

Anomaly Detection and Error Correction

Automation for Anomaly Detection & Correction

Manual recovery where necessary

Introduction of Anomaly Detection – But Dependency Remains
The Expanding Partnership

Coronavirus Resource Center

JHU CSSE COVID-19 Dashboard

Johns Hopkins Bloomberg School of Public Health

Centers for Civic Impact

Johns Hopkins Whiting School of Engineering

Center for Systems Science and Engineering

Johns Hopkins Sheridan Libraries

APL

Johns Hopkins Applied Physics Laboratory

Esri

THE SCIENCE OF WHERE™
Gen 2 Architecture

Expanded Data Sourcing

Autonomous Collection

"Authoritative" Data Sources

Global Sitreps

National Reporting

State & Local Reporting

Crowd Sourced Data

~400

Collection

Clean Data Store

Data Cleaning Process

Raw Data Store

Data Fusion Process

End Product

Anomaly Detection & Recovery

Open Data Sharing

Automated Updates

"Authoritative" Data Sources

Autonomous Collection

Global Sitreps

National Reporting

State & Local Reporting

Crowd Sourced Data

~400

Collection

Clean Data Store

Data Cleaning Process

Raw Data Store

Data Fusion Process

End Product

Anomaly Detection & Recovery

Open Data Sharing

Automated Updates
Every morning, the first website I consult is the Johns Hopkins University (JHU) COVID-19 dashboard. It is the go-to resource to track the global pandemic. Public-health authorities use it to guide policies, with the dashboard featured prominently in emergency operation centers around the world. Researchers depend on it for their analysis and modeling. Virtually every news organization uses the JHU data as the basis for their reporting.

- Dr. Leana Wen
What’s Next

A need for open public data standards and sharing in light of COVID-19

The disjointed public health response to the COVID-19 pandemic has demonstrated one clear truth: the value of timely, publicly available, high-quality data is imperative for systematically collecting, visualizing, and sharing information on emerging infectious and notifiable diseases in real-time.

Moving forward, it is imperative that a standardized reporting system for systematically collecting, visualizing, and sharing high-quality data on emerging infectious and notifiable diseases in real-time is established. The data should be made available at a spatial and temporal scale that is granular enough to prove useful for planning and modelling purposes. Additionally, a critical component of the proposed system is the democratization of data; all collected information (observing necessary privacy standards) should be made publicly available immediately upon release, in machine-readable formats, and based on open data standards.
Acknowledgements