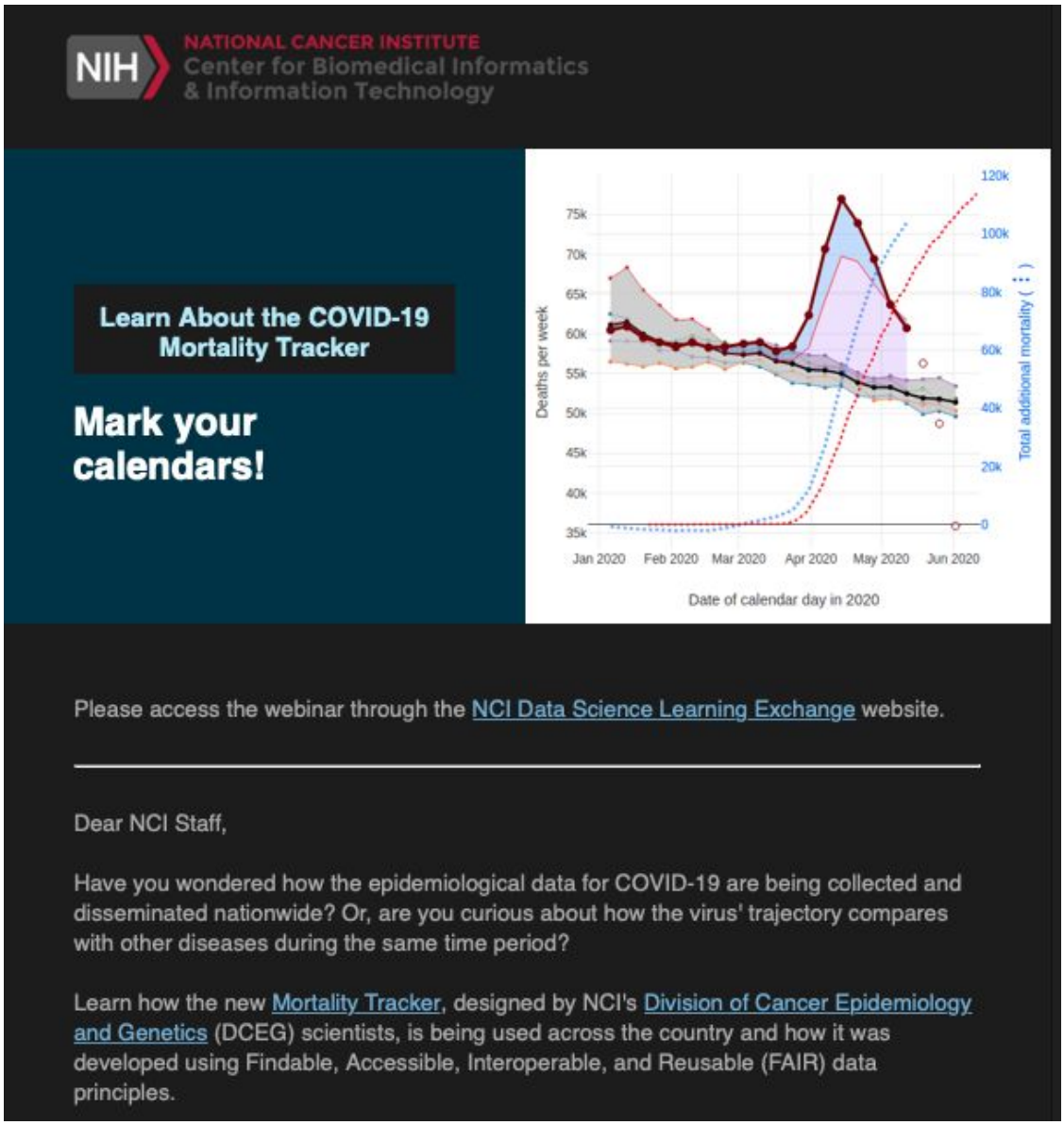
bit.ly/mtrack-2020-06-24



NCI Division of Cancer Epidemiology and Genetics (DCEG), NCI Data Science Learning Exchange. → being recorded, slides/notes public

- Jonas S. Almeida, Ph.D., Chief Data Scientist, (*)
- Amy Berrington, PhD, Radiation Epidemiology Branch Chief
- Neal Freedman, PhD, Senior Investigator
- Meredith Shiels, PhD, Investigator
- Praphulla Bhawsar, MS, Data Engineer (*)
- Bhaumik Patel, MS, Software Engineer
- Montserrat Garcia-Closas, MD PhD, Integrative Tumor Epidemiology Branch Chief

* open mic

bit.ly/mtrack-2020-06-24

Mortality Tracker

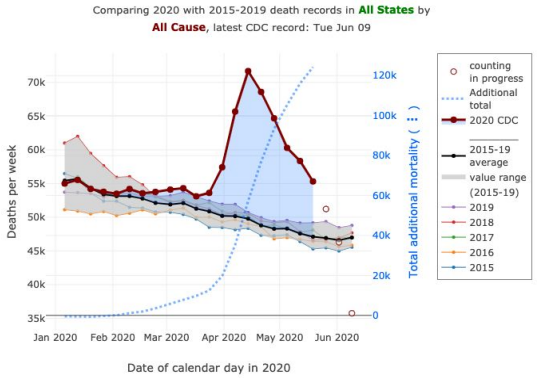
A demonstration of the FAIR implementation of a real-time mortality tracking tool and a discussion of what Data Commons in the age of COVID.

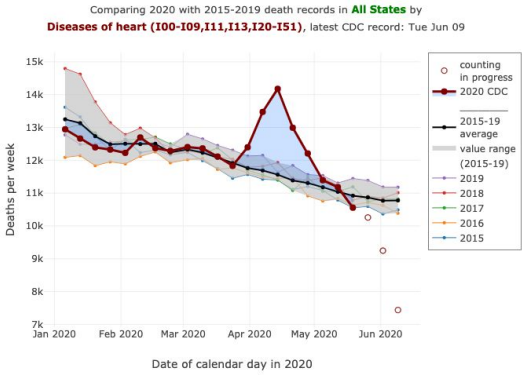
No-downloads no-pay design

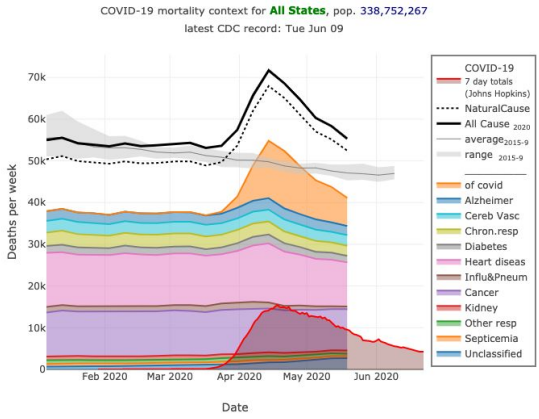
Guilty as charged: the tracker tool indeed uses other people's data and other people's computational resources. Which is just what Data Commons are about. Can we do it in real time?

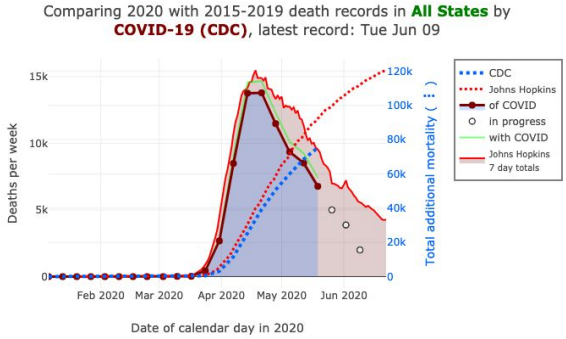
1. Demonstration of the tool

(~15 min)
episphere.github.io/mortalitytracker,
bit.ly/mortalitytracker









2. A disintermediated Architecture

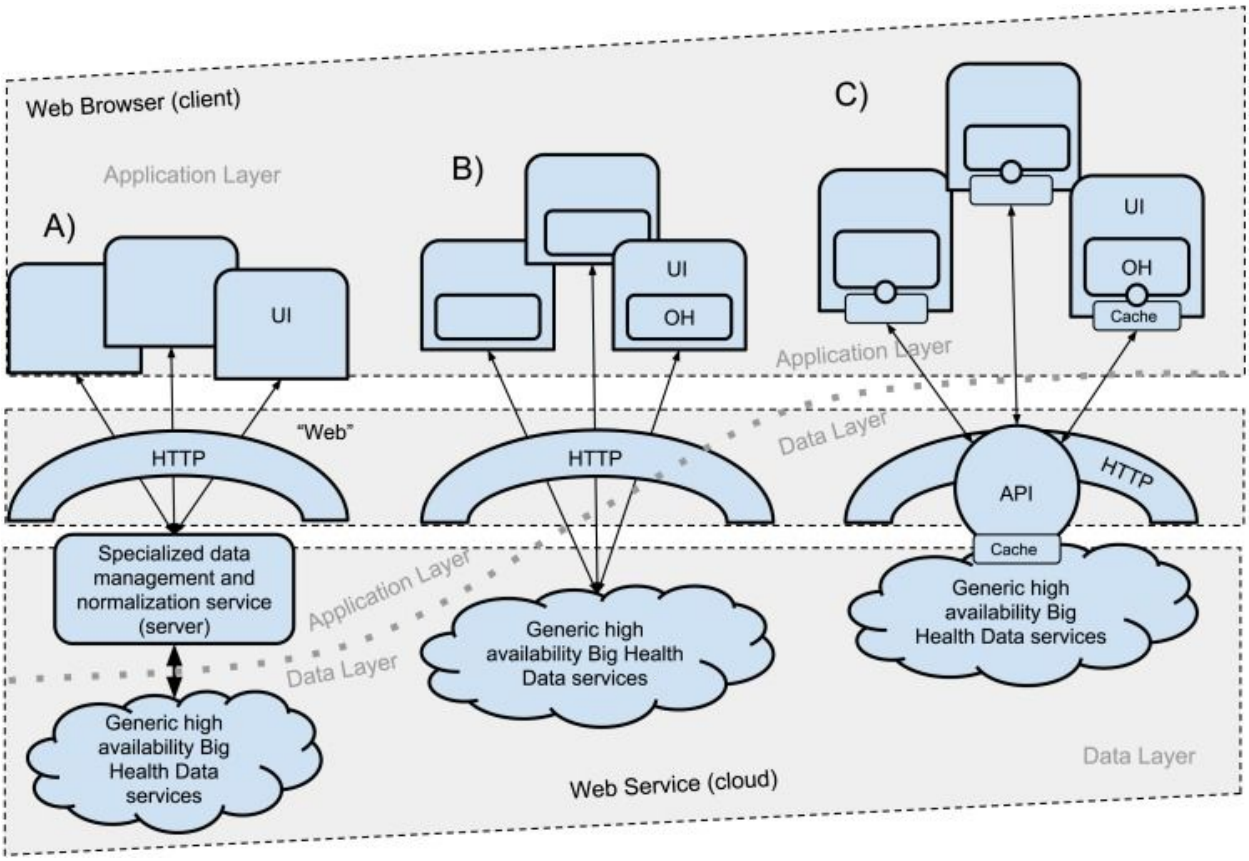


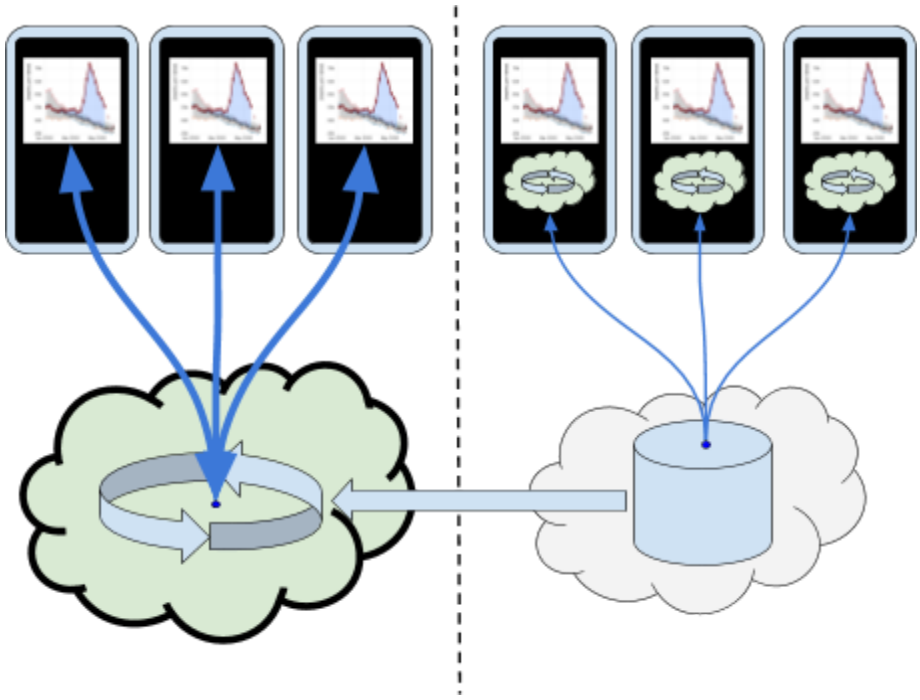
Figure 1. Evolving Web Computing Architectures. Evolution of the API economy from its pre-REST stage (A) to stateless transfer via HTTP (B), recently abstracted by constructs like GraphQL that combine an API language with a query engine (C). The prototype accompanying this report uses SoQL (see 'Methods') to illustrate the viability of the latter design, where the traversal of the Data Layer is abstracted as a stateless backend. The Cloud instantiation of this model approaches the description of BaaS (Backend-as-a-service).

3. Preceding work - bit.ly/loadsparcs

Almeida JS, J Hajagos, J Saltz J, M Saltz (2019) **Serverless OpenHealth at data commons scale—traversing the 20 million patient records of New York's SPARCS dataset in real-time**. PeerJ 7:e6230 [PMID:30671301].

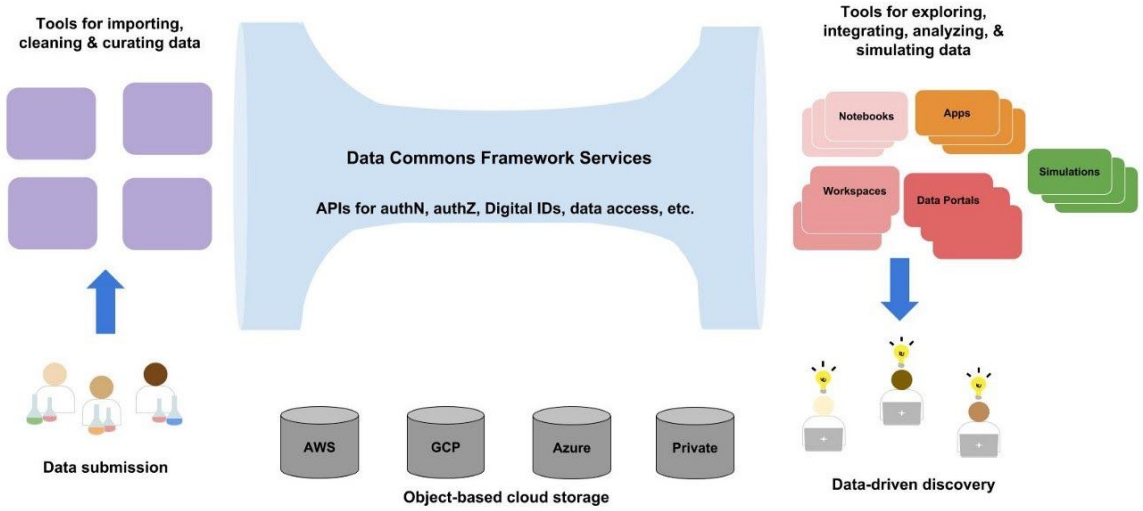
4. The implications of lacking a middle layer

It has to be somewhere else …
How do others do it - the Media for example - the NYTimes story.
How do others enable it - COVID data at Johns Hopkins for example
Cases-studies: dependency on ArcGis at NIEHS[1] and Johns Hopkins.
… or not, as in https://www.cdc.gov/covid-data-tracker/index.html
CDC is also doing it - https://www.cdc.gov/covid-data-tracker/index.html



[1] Thank you, Alison Motsinger-Reif (NIH/NIEHS)

5. Data Commons - discussion



@ Bob Grossman "A Proposed End-To-End Principle for Data Commons" 2018

A. Let's start with the iconic data commons resource - NCI's Genomic Data Commons.
B. How will it scale across omics? - https://datascience.cancer.gov/data-commons.
C. Discuss - what does COVID teach us about the predictability of API's?
D. "Don't force me into your walled garden, I've been cultivating my own"

The API ecosystem in the narrow middle is a market for demonetized data services.

6. Web APIs - Data Economy

Distributed data aggregation in real time creates its own data economy

A. Demonstration with Mortality tracker: where does data wrangling take place …
B. CDC changes data structures and variable names like everybody else …

How do people pay

- by advancing demonetized data assets. For example, NYT public data assets compete with official sources, just like wikipedia does … We can use the same literate programming model.

How do people trade

- they trade in persistent resources (for example, what is most perishable in https://mathbiol.github.io/tcgatil ?).

7. The Technology landscape moves fastest

The Technology landscape moves faster than we can.

A. Only 10% of IT is in the Cloud, but most will be there eventually [The Economist]
B. By 2023 most cloud computing will take place at the edge [Gartner Report]
   Edge computing Complements and extends Cloud computing

8. Feedback, announcements

Everybody please: bit.ly/DCEG_Tracker_Feedback

Cloud4Bio: back to gotomeeting for AI Federated Learning (Jeya) ...