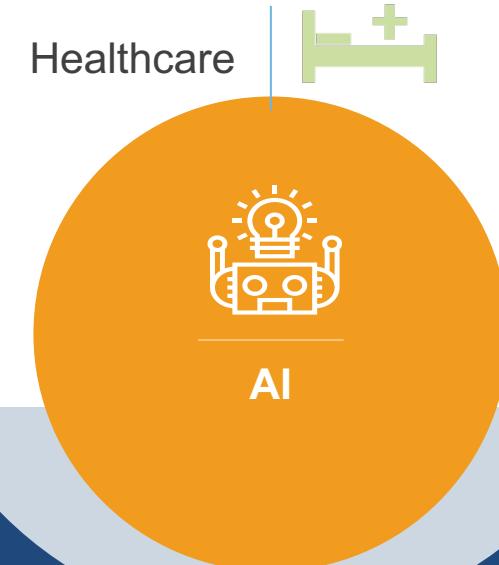


COVID-19 Data Index – Making datasets findable and accessible



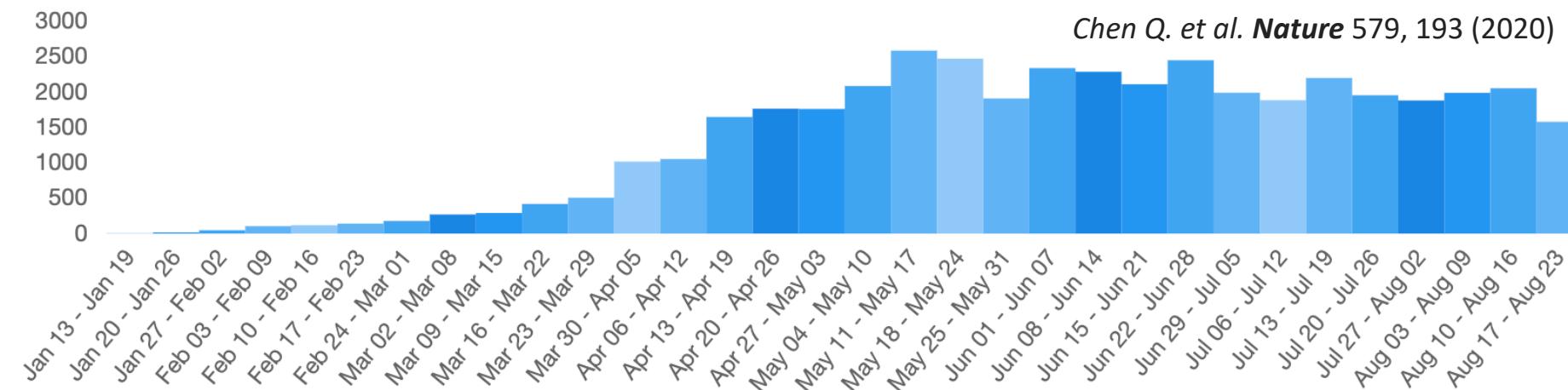
Hua Xu PhD
School of Biomedical Informatics, University of Texas, Houston

August 24th, 2020
2020 KDD Workshop on Applied Data Science for Healthcare

Big Data Generated from COVID-19 Research

- Large volume of publications: 43,076 (LitCovid, 8/23)

Weekly Publications



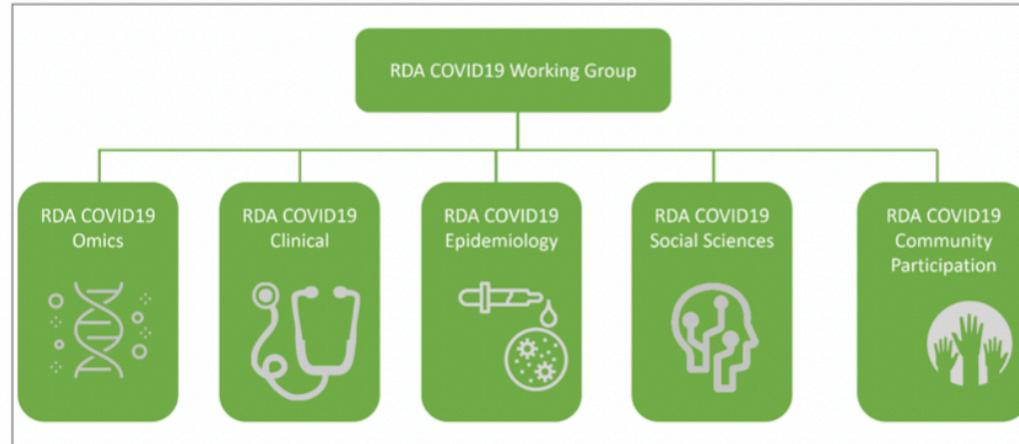
- Heterogenous types of datasets

- Epidemiological data on testing and case statistics at various locations
- Omics data from labs
- Clinical data from surveys, studies (e.g., imaging, assays) or from electronic health records
- Administrative (e.g., PPE, ventilators, hospitalizations, ICU beds)
- Socio-demographic, environmental, economic, individual mobility and transportation data.

Effort on Sharing COVID-19 Data

■ FAIR Principles (Findable, Accessible, Interoperable, and Reusable)

The screenshot shows the VODAN website under the GO FAIR banner. The main heading is "Virus Outbreak Data Network (VODAN)". Below it, there's a circular logo with "DATA TOGETHER" in the center, surrounded by "CODATA", "GO FAIR", "RDA", and "WORLD DATA SYSTEM". The page content discusses the spread of the COVID-19 virus and the challenges of managing and reusing data from past epidemics like Ebola.



■ Other data sharing initiatives

The screenshot shows the CORD-19 website. The main heading is "CORD-19 COVID-19 Open Research Dataset". Below it, there's a "Get Started" button. The page content discusses the Semantic Scholar team's partnership with leading research groups to provide CORD-19, a free resource of more than 130,000 scholarly articles about the novel coronavirus.

The screenshot shows an RSNA News article titled "RSNA Announces COVID-19 Imaging Data Repository". The article states: "Planned open data repository will be for international COVID-19 imaging research and education efforts". The date "March 31, 2020" is at the bottom.

The screenshot shows the NIH Office of Data Science Strategy COVID-19 page. The main message is: "COVID-19 is an emerging, rapidly evolving situation." It includes links to CDC, NIH, and other federal agencies' resources. The page also mentions "Open-Access Data and Computational Resources to Address COVID-19".

Open-Access Data and Computational Resources to Address COVID-19

COVID-19 open-access data and computational resources are being provided by federal agencies, including NIH, public consortia, freely available to researchers, and this page will be updated as more information becomes available.

Our Approach – COVID-19 Data Index (www.covid19dataindex.org)

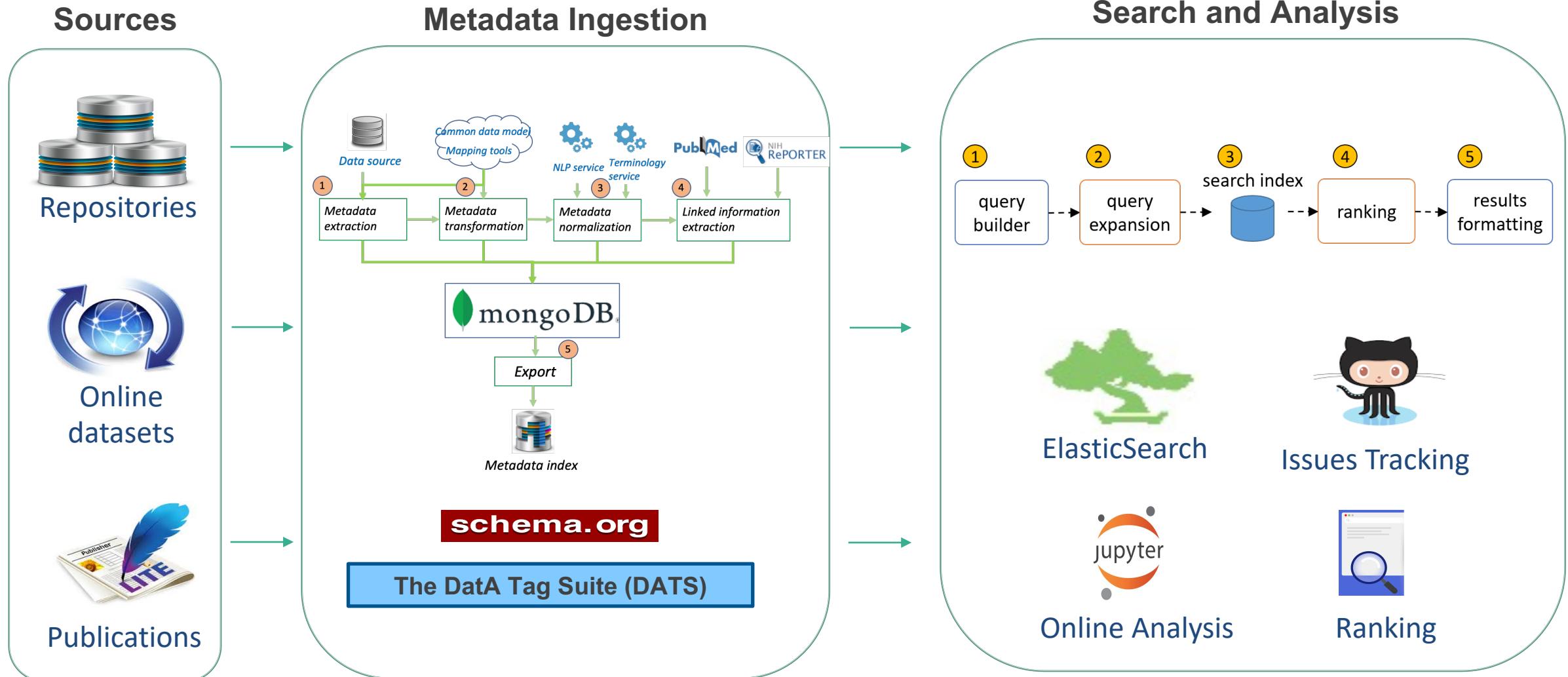
The screenshot shows the COVID-19 Data Index homepage. On the left, two boxes highlight features: 'Diverse Data Types' points to the 'Data Type' section which lists various categories like Clinical, Imaging, Epidemiology, etc., with counts such as 1853 for Clinical and 1108 for Imaging. 'Different Repositories' points to the 'Repository' section which lists platforms like GitHub, ClinicalTrials.gov, Kaggle, etc., with counts such as 2712 for GitHub and 1730 for ClinicalTrials.gov. The main content area includes an overview section with three charts: a bar chart showing 'Data Sets : 6017' from March to July, a pie chart showing 'Data Types : 10' (Clinical, Imaging, Epidemiology, Social Science, Rest), and a bar chart showing 'Countries : 118' (Multi-country, United States, India, France, China). Below this are sections for 'Popular Datasets' featuring 'Coronavirus COVID-19 Global Cases by the Center for System Science and Engineering (CSSE) at Johns Hopkins University (JHU)' and 'Nextstrain-Real time tracking of pathogen evolution'.

Title :	Coronavirus COVID-19 Global Cases by the Center for System Science and Engineering (CSSE) at Johns Hopkins University (JHU)
Description :	This is the data repository for the 2019 Novel Coronavirus Visual Dashboard operated by the Johns Hopkins University Center for Systems Science and Engineering (JHU CSSE). Also, Supported by ESRI Living Atlas Team and the Johns Hopkins University Applied Physics Lab (JHU APL).
Organization :	the Johns Hopkins University Center for Systems Science and Engineering (JHU CSSE)
Source Url :	 https://www.arcgis.com/apps/opsdashboard/index.html#/bda7594740fd40299423467b48e9ecf6
Download data :	 https://github.com/CSSEGISandData/COVID-19/archive
Publication Name :	An interactive web-based dashboard to track COVID-19 in real time

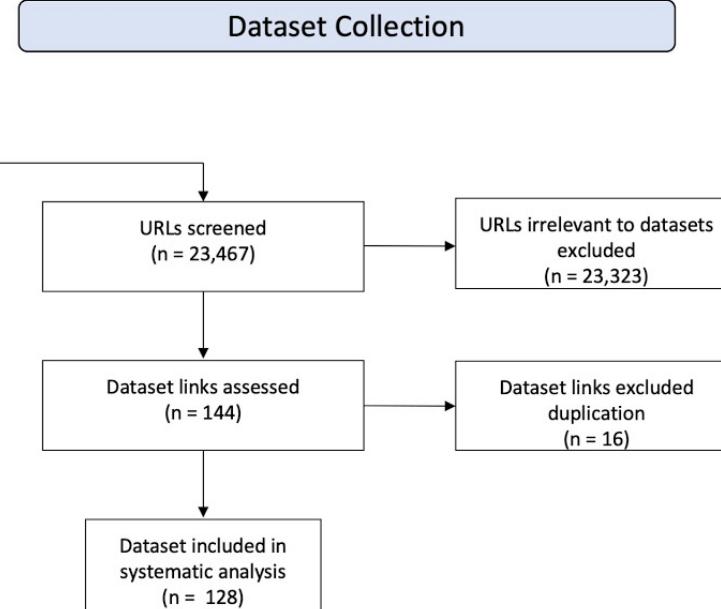
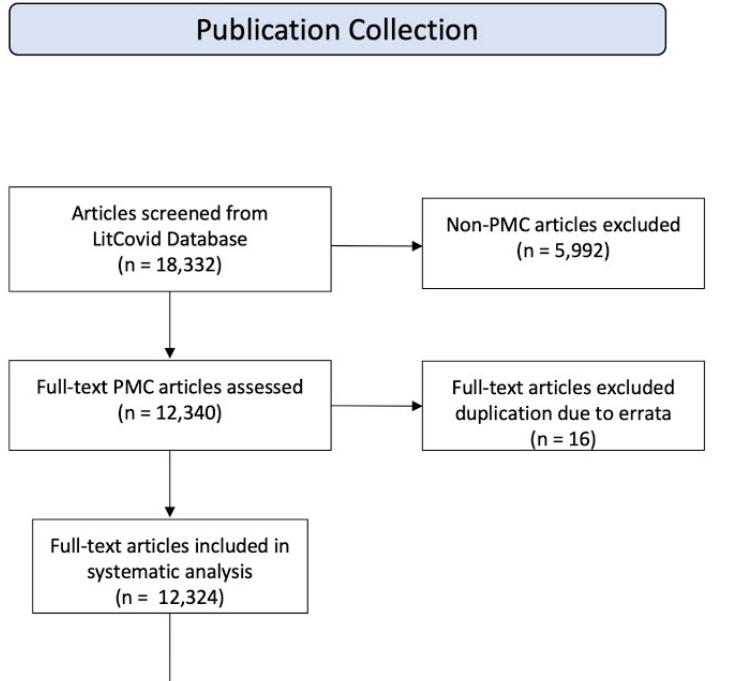
Links to detailed information and direct downloading

Ohno-Machado and Xu. *Nature* 584, 192 (2020)

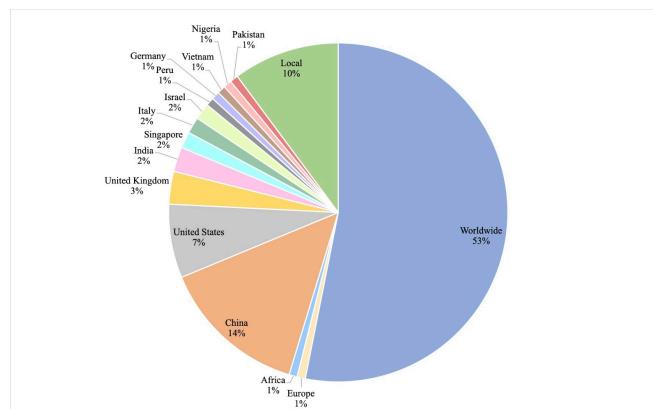
COVID-19 Data Index Workflow



A review of COVID-19 datasets in PubMed Central articles



- Research questions
 - Accessibility
 - Content
 - Citation



Most Cited COVID-19 Datasets in PMC

Dataset	Overall Citations	URL Citations	Article Citations
John Hopkins University Dashboard ²	454	416	275
Real-time estimation of the novel coronavirus incubation time ³	239	0	239
Worldometers ³⁴	231	231	0
Substantial undocumented infection facilitates the rapid dissemination of novel coronavirus (SARS-CoV-2) ³⁷	189	0	189
Estimates of the severity of coronavirus disease 2019: a model-based analysis ³⁸	132	0	132
Feasibility of controlling COVID-19 outbreaks by isolation of cases and contacts ⁵	104	0	104
Pattern of early human-to-human transmission of Wuhan 2019 novel coronavirus ³⁹	102	1	102
Early dynamics of transmission and control of COVID-19: a mathematical modelling study ³	97	0	97
The effect of travel restrictions on the spread of the 2019 novel coronavirus (COVID-19) outbreak ⁶	90	0	90
CDC ³⁵	87	87	0

Summary

- Data sharing is critical for reproducible research including COVID-19
- FAIR Principles are important for sharing digital assets; but significant efforts are needed to develop FAIR-compliant datasets
- Citation analysis could be the incentive for data sharing; but more work is need for formally citing datasets and measuring its impact



Thank you!
Questions?

hua.xu@uth.tmc.edu