AWS launches machine learning enabled search capabilities for COVID-19 dataset

by Taha A. Kass-Hout, MD, MS and Ben Snively | on 20 APR 2020 | in [Amazon Comprehend Medical](https://aws.amazon.com/blogs/publicsector/category/artificial-intelligence/amazon-comprehend-medical/), [Amazon Kendra](https://aws.amazon.com/blogs/publicsector/category/artificial-intelligence/amazon-kendra/), [Amazon Machine Learning](https://aws.amazon.com/blogs/publicsector/category/artificial-intelligence/amazon-machine-learning/), [Announcements](https://aws.amazon.com/blogs/publicsector/category/post-types/announcements/" \o "View all posts in Announcements), [Artificial Intelligence](https://aws.amazon.com/blogs/publicsector/category/artificial-intelligence/" \o "View all posts in Artificial Intelligence), [Education](https://aws.amazon.com/blogs/publicsector/category/public-sector/education/), [Government](https://aws.amazon.com/blogs/publicsector/category/public-sector/government/), [Healthcare](https://aws.amazon.com/blogs/publicsector/category/industries/healthcare/), [Healthcare](https://aws.amazon.com/blogs/publicsector/category/public-sector/healthcare-public-sector/), [Nonprofit](https://aws.amazon.com/blogs/publicsector/category/public-sector/nonprofit/), [Public Sector](https://aws.amazon.com/blogs/publicsector/category/public-sector/) | [Permalink](https://aws.amazon.com/blogs/publicsector/aws-launches-machine-learning-enabled-search-capabilities-covid-19-dataset/) | [Share](https://aws.amazon.com/fr/blogs/publicsector/aws-launches-machine-learning-enabled-search-capabilities-covid-19-dataset/)

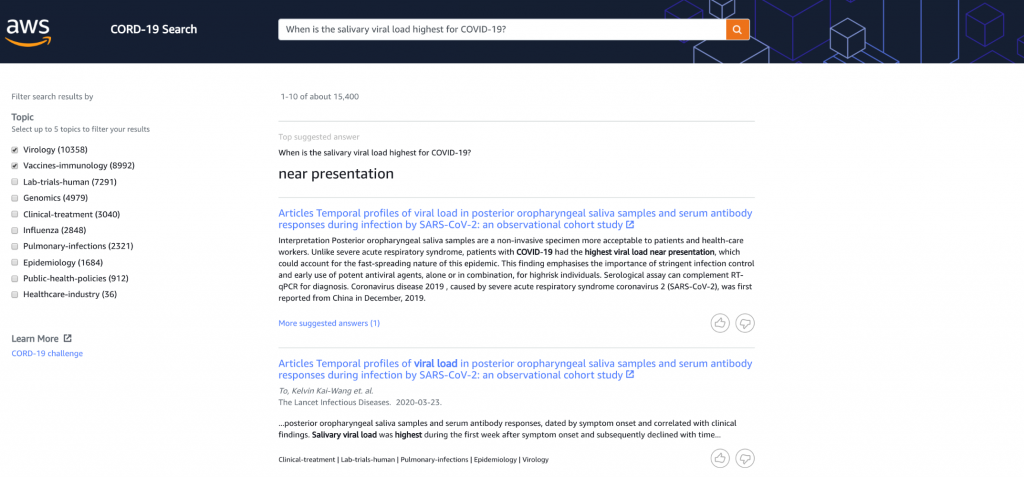
As the world grapples with COVID-19, researchers and scientists are united in an effort to understand the disease and find ways to detect and treat infections as quickly as possible. Today, Amazon Web Services (AWS) launched [CORD-19 Search](https://cord19.aws/), a new search website powered by machine learning that can help researchers quickly and easily search tens of thousands of research papers and documents using natural language questions.

As part of the White House remote roundtable with the tech sector held last month, the Allen Institute for AI (AI2) released CORD-19 (COVID-19 Open Research Dataset). CORD-19 Search was built leveraging this [dataset](https://registry.opendata.aws/cord-19/" \t "_blank), which initially consisted of approximately 24,000 scientific and research sources related to COVID-19, SARS-CoV-2, and coronaviruses. Since it was made available, the CORD-19 dataset has nearly doubled to 47,000 research papers and documents sourced from peer-reviewed publications and pre-print servers.

The scientific community is responding to the threat of COVID-19 by studying the novel coronavirus and publishing cutting-edge research and findings on detection and treatment. This body of work is generating scientific and medical evidence on COVID-19 at an exponential scale – so much so, that it is difficult to digest and analyze. Making key insights within such a large amount of information discoverable is critical to developing responses to disease transmission and treatment, including finding a cure or vaccine for COVID-19.

CORD-19 Search helps researchers navigate this fast-growing body of coronavirus literature to efficiently find relevant and up-to-date information. CORD-19 Search provides a simple search interface where researchers can ask questions using natural language such as, “When is the salivary viral load highest for COVID-19?” and “Is convalescent plasma therapy a precursor to vaccine?” CORD-19 Search produces precise answers as well as source documents.

For example, the answer to COVID-19’s highest viral load states that, “Salivary viral load was highest during the first week after symptom onset and subsequently declined with time.” Similarly, CORD-19 Search responds that convalescent plasma therapies, “in the absence of vaccine would provide a stopgap measure, ideally consider to give to those who are at risk of exposure or early in showing symptoms (as a preparedness measure)” along with related scientific articles from past trials during SARS and Ebola. CORD-19 Search also provides evidence-based topics on incubation, transmission, therapeutics, and risk factors. This functionality is of enormous value to scientists who can quickly query, validate their research, and advance their investigations.

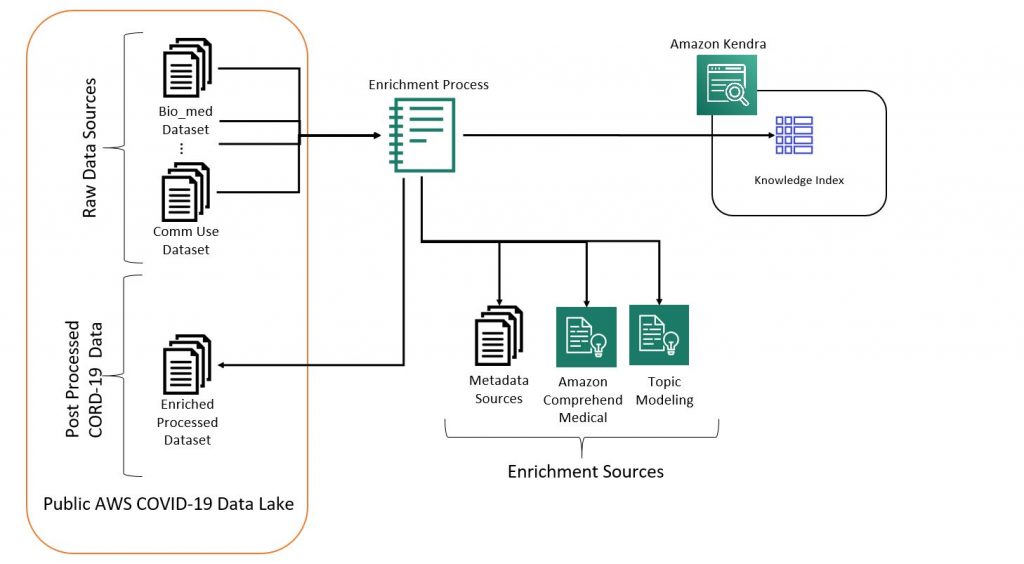
[](https://d2908q01vomqb2.cloudfront.net/9e6a55b6b4563e652a23be9d623ca5055c356940/2020/04/20/Sample-results-from-CORD-19-Search.png)

*Sample results from CORD-19 Search*

**How AWS built CORD-19 Search**

CORD-19 Search uses AWS machine learning services to power comprehensive and actionable results. The original dataset is enriched with [Amazon Comprehend Medical](https://aws.amazon.com/blogs/machine-learning/introducing-medical-language-processing-with-amazon-comprehend-medical/), a natural language processing service that uses machine learning to extract relevant medical information from unstructured text, including disease, treatment, and timeline. The data is then mapped to clinical models and medical topics associated with COVID-19 using a [multi-label classification model](https://docs.aws.amazon.com/comprehend/latest/dg/how-document-classification.html#multiclass-multilabel2) and [inference](https://docs.aws.amazon.com/comprehend/latest/dg/extracted-med-info-V2.html" \t "_blank), such as virology, immunology, and laboratory or clinical trials. The information is then indexed in [Amazon Kendra](https://aws.amazon.com/kendra), a highly accurate enterprise search service powered by machine learning, delivering robust natural-language query capabilities that make it easier to find and rank related articles. The Amazon Comprehend Medical enriched data and Amazon Kendra search are built from data available in the [public AWS COVID-19 data lake](https://covid19-lake.s3.us-east-2.amazonaws.com/dashboard.html?dashboardName=COVID-19), where anyone can experiment with and analyze curated data related to the disease and share their results.

“One of the most immediate and impactful applications of AI is in the ability to help scientists, academics, and technologists find the right information in a sea of scientific literature to move research faster. The Allen Institute for AI, and particularly the Semantic Scholar team, is committed to providing this important resource and supporting the associated AI methods the community is using to tackle this crucial problem.” *– Dr. Oren Etzioni, Chief Executive Officer of the Allen Institute for AI*

[](https://d2908q01vomqb2.cloudfront.net/9e6a55b6b4563e652a23be9d623ca5055c356940/2020/04/20/CORD-19.jpg)

*CORD-19 Search architecture*

**The long-term benefits of CORD-19 Search**

AWS is applying machine learning to the CORD-19 data set to accelerate the pace of discovery, where the speed of COVID-19 disease intervention, progression, and treatment is critical. Our long-term vision is to build future capabilities based on the CORD-19 Search architecture to integrate disparate data sources, including clinical research data, to allow researchers around the world to aggregate patient-specific patterns of disease progression, provide data-driven decisions, and positively impact patient outcomes at scale.

We are committed to serving the scientific community and general public to support the global response to COVID-19. CORD-19 Search is now publicly available at [https://cord19.aws.](https://cord19.aws/)

TAGS: [Amazon Comprehend Medical](https://aws.amazon.com/blogs/publicsector/tag/amazon-comprehend-medical/), [Amazon Kendra](https://aws.amazon.com/blogs/publicsector/tag/amazon-kendra/), [Artificial Intelligence](https://aws.amazon.com/blogs/publicsector/tag/artificial-intelligence/), [COVID-1](https://aws.amazon.com/blogs/publicsector/tag/covid-19/)