Help Spread The Word

Overview
- Team
- Project description
  - Motivation and proof of a problem existing
- Project Impact
- Work to date
  - Proof of Concept
  - CoronaWhy Data Lake
  - CoronaWhy Common Research and Data Infrastructure
  - CoronaWhy Services
  - Workshops and Public Talks
  - Published Studies
  - Demo
  - Feedback from Researchers
- Target completion
- Use of funds
- Other funding

Team Highlights
- Janosch Ortmann - Assistant Professor in Data Science and Business Intelligence, ESG UQÀM
- Dan Sosa - Ph.D. Candidate in Biomedical Informatics at Stanford University
- Christine Chen, MD MPH - PhD Candidate & Assistant Policy Researcher at RAND Corporation | Data Scientist | Team Lead at CoronaWhy
- Dr Tayab Waseem - Public Policy Fellow for the AAI; Director of Medical Informatics and AI Integration at WMRC
- Vyacheslav Tykhonov, Senior Information Scientist at DANS-KNAW | Director of CoronaWhy Data Infrastructure and Services
- Artur Kiulian - Serial Entrepreneur, Technologist

Advisors
- Dr. Byron Wallace (The IRB for the utility study is being run through him at North Eastern University)
- Benjamin M. Gyori - Research Associate in Therapeutic Science at Harvard Medical School
- Serhii Myroshnychenko - MS, Epidemiology & Biostatistics, Pharma Marketing MBA Fellow
- Jack Park - AI Researcher and Open-Source Advocate
- Emad Mostaque - CAIAC & HAI

Jeremy Zucker - Computational Biologist at Pacific Northwest National Laboratory
Charles Tapley Hoyt - Computational Biologist at Enveda Therapeutics

Help Spread The Word

Tweet about our application here (tell the world and Elon Musk about us), share on Facebook or share on Linkedin.

Overview

Current document is a proposal for funding for the continuation of development of an **Open Source Artificial Intelligence (AI)-powered Literature Review Discovery Engine** for efficient review and navigation of medical literature focused on the COVID-19 pandemic.

Team

Janosch Ortmann - PhD, PI
Affiliation: ESG UQAM (École des sciences de la gestion)
Role: Assistant Professor
Year received: Phd received in 2013
Your top three published papers:

- https://www.frontiersin.org/articles/10.3389/fnins.2020.00207/full?report=reader
- https://www.cirrelt.ca/documentstravail/cirrelt-2018-39.pdf
- https://projecteuclid.org/euclid.ejp/1465067131

Dan Sosa. PhD student, Lead Researcher
Christine Chen. MD, MPH & PhD Candidate, Lead Researcher for Search Engine
Tayab Waseem. PhD, Director of Medical Informatics & AI integration
Vyacheslav Tykhonov, MSCS, Director of Data Infrastructure and Services
Anton Polishko, PhD, Director of Cloud Infrastructure
Artur Kiulian. MSAI, Director of Product
Mayya Lihovodov - Head of Partnerships
Serhii Myroshnychenko MS, MBA - Head of Product
Bianca Grizhar - MSAI, Lead Architect
Mike Honey - Technical Lead, Data Visualization
Tyler Parker-Smith - Product Lead
Nicole Macam - UX researcher
Yuan Li - UI designer
Katharine Miller - Biomedical Writer / Advisor
Svetlana Tchistiakova - NLP researcher & engineer
[Web developers]
[Marketers/Communicators]

[Medical annotators]

Project description

Motivation and proof of a problem existing

The Covid-19 research community, healthcare workers and policymakers are all struggling to remain informed as thousands of new scientific papers are released weekly. The Covid-19 Open Research Dataset (CORD-19), already consists of more than 181,000 Covid-19-related scholarly articles (including more than 60,000 with full text). CORD-19 is the first large scale COVID-19 research data set of its kind within the United States, and is spearheaded by the White House Office of Science and Technology Policy.

This consortium is the first of its kind, combining experts in the fields of biology, epidemiology, medicine, healthcare research, bioinformatics and artificial intelligence (AI) into one large cross-disciplinary consortium. Partners of CoronaWhy include the **Allen Institute for AI**, **Chan Zuckerberg Initiative (CZI)**, **Georgetown University's Center for Security and Emerging Technology (CSET)**, **Microsoft**, and the National Library of Medicine (NLM) at the National Institutes of Health (NIH). Academic Partners include scientists and clinicians from over 25 medical institutes. The team uses advanced natural language processing to expedite the analysis of the increasing amounts of COVID-19 related research in real time to aid in providing a centralized platform for structured query searches within the COVID-19 scientific literature.

With this proposal, we request support to continue research and development of an innovative and potentially transformative **Artificial Intelligence (AI) driven "Literature Review Discovery Engine"** for efficient review and navigation of epidemiological and clinical research related literature focused on the COVID-19 pandemic.

The effort will be undertaken in two phases:

1. Strengthen and extend our existing scientific research for the curation of data, to include a larger set of research questions validated by epidemiologists used as the basis for data discovery with the publicly available CORD-19 dataset [1];

2. Further research, develop and test the required NLP and ML models for an automated AI discovery engine that parses user inputs directly to produce optimized searches, both on CORD-19 and with the addition of broader data samples from CoronaWhy's Distributed Open Data Infrastructure [2].

These tools satisfy a growing need among communities of researchers and medical professionals seeking to parse the rapidly evolving corpus of information on COVID-19. This need is becoming increasingly acute as information about the virus is generated in

unprecedented volumes and with great urgency worldwide. There is a growing evidence of a gap that has been studied across different disciplines [3]

"The COVID-19 literature has grown in much the same way as the disease's transmission: exponentially. The NIH's COVID-19 Portfolio, a website that tracks papers related to the SARS-CoV-2 coronavirus and the disease it causes, lists more than 28,000 articles — far too many for any researcher to read. But a fast-growing set of artificial-intelligence (AI) tools might help researchers and clinicians to quickly sift through the literature. © Artificial-intelligence tools aim to tame the coronavirus literature

https://www.nature.com/articles/d41586-020-01733-7

Our approach uses Natural Language Processing (NLP) and Machine Learning (ML) to parse human-readable questions into queries that are used for extracting quantitative and qualitative information from the corpus of literature, extending the scope of the traditional search engine to allow for data extraction of enhanced complexity and scope.

"Advancements in computation and machine learning have enabled natural language processing (NLP) techniques that are effective and scalable for processing large bodies of unstructured text. Recently, NLP was applied to all ~28.6 million PubMed abstracts to synthesize and summarize the relationships between drugs, genes/proteins, and diseases into a heterogeneous knowledge graph known as the Global Network of Biomedical Relationships (GNBR). © A Literature-Based Knowledge Graph Embedding Method for Identifying Drug Repurposing Opportunities in Rare Diseases

Read more here:

https://drive.google.com/file/d/1xDE1LutuDgo4HOtH8IzkIHEIvFspqS9w/view?usp=sharing

Project Impact

CoronaWhy is a globally distributed research initiative, established in March 2020 and driven entirely by over 1500 volunteers to date. With the support of FastGrants we are confident that the technical skills of contributors can be used to make a meaningful and more importantly **immediate impact** against COVID-19, delivering new tools to facilitate researchers worldwide in their ongoing efforts to understand and mitigate the effects of this unprecedented pandemic, while potentially opening a new paradigm for mass data discovery in large datasets.

Work to date

Proof of Concept

https://www.coronawhy.org/literature-review-demo

CoronaWhy Data Lake

The vision of the CoronaWhy community is to build an Artificial Intelligence infrastructure completely from Open Source components and with publicly available Machine Learning models

developed by various organizations worldwide. All data curated and published using the Dataverse data platform developed by Harvard IQSS and provided by CoronaWhy as a service that allows to keep provenance information for every single data file.

https://datasets.coronawhy.org

CoronaWhy Common Research and Data Infrastructure

At CoronaWhy we are building a Common Research and Data Infrastructure for Open Science that can be used by researchers coming from various scientific communities involved in COVID-19 research. This distributed and scaled Cloud infrastructure follows Reproducible Science and FAIR principles and should be suitable for other important scientific challenges such as cancer and AIDS research. It can be also installed locally and exposed as a number of services.

http://github.com/CoronaWhy/coronawhy-infrastructure

CoronaWhy Services

CoronaWhy organization provides Common Services like Biological Expression Language (BEL), Hypothes.is annotation tool and CoronaWhy Colab to support other COVID-19 vertical projects, not necessary with CoronaWhy affiliation.

The ultimate goal is to build Medical, Social and Economic Knowledge Graphs to make it possible to understand not only Biological and Environmental Factors of COVID-19 spread but the Social and Economic impacts as well, and investigate how quarantine and social distancing measures affected the population. The access to all available knowledge graphs delivered by appropriate CoronaWhy services like Virtuoso with SPARQL endpoint and exposed as Linked Data.

https://www.coronawhy.org/services

Workshops and Public Talks

- Building an AI-powered Literature Review for COVID-19 at SciNLP workshop: “Natural Language Processing and Data Mining for Scientific Text” organized by Allen Institute for AI and Google Research on June 27, 2020. The abstract and video introduction available online on SciNLP website.
- Fighting COVID-19 through Data - Professor Agnis Stibe, founder of the Transforming Wellbeing Theory (TWT), April 16, 2020
- CoronaWhy: Re-curation and rational enrichment of knowledge graphs in BEL presented by Charles Tapley Hoyt (https://cthoyt.com), May 6, 2020
- Fight against COVID-19 - DANS presentation by Vyacheslav Tykhonov, June 3, 2020
- Charles Tapley Hoyt presented a workshop Writing Reusable, Reproducible Python: Documentation, Packaging, Continuous Integration, and Beyond, June 3, 2020

- Global Information Technology Management Association (GITMA) Conference - Managing the Chaos - Transformative Power of Flat Decentralized Communities, June 22, 2020
- CoronaWhy Common Research and Data Infrastructure at SciNLP workshop: “Natural Language Processing and Data Mining for Scientific Text” organized by Allen Institute for AI and Google Research on June 27, 2020. The abstract and video introduction available online on SciNLP website.

Published Studies

IRB approved utility study of an AI-powered literature review is currently under revision and set to be published in August of 2020, authored by team member Tayab Waseem and 250+ participating medical researchers. Validation studies will be published later in 2020 after successful creation and adoption of the proposed software.

"The utility of AI powered literature reviews in rapidly generating answers to critical COVID-19 questions" - expected submission to NEJM by August 1st, 2020

Demo

https://www.loom.com/share/a645bfba65bc43ab8c11590a0300fbba

Feedback from Researchers

What feedback have you had from practitioners in the field so far?

Currently our team has over 150 medical and scientific volunteers across 30 institutes working on creating an AI-driven live literature review tool. We asked a couple of our team members across specialties, ranging from medical students, residents, and attending physicians to PhD students and postdocs how this project will impact the future of their field. In their own words:

***Jose Morey (M.D., Eisenhower Fellow, Chief Medical Innovation Officer - Liberty BioSecurity)**: Using AI to augment literature reviews is the epitome of man + machine moving the needle forward for medicine. It is a perfect example of human-centered AI and its best.*

***Maikel Boot, PhD (Postdoctoral Fellow at Yale University)**: Having an AI/ML search engine provide a landscape analysis with relevant metrics of any given research topic would be a game-changer for writing grants, chapters and reviews. Especially when fields are rapidly developing or have large bodies of literature, having a detailed overview of all relevant literature saves a lot of time.*

***Michael Stolz, M.D. (Surgery Resident at Northeast Georgia Medical Center)**: Increase the efficiency of time spent looking for answers will allow me to spend more time caring for our patients using the latest and most relevant data available.*

Lucas Buyon (PhD candidate at Harvard)*: The hope is the AI-powered approaches will dramatically reduce the time spent writing literature reviews. Furthermore, as the rate of scientific publication continues to increase, AI approaches may allow for the publication frequency of subject reviews to better keep pace the ever-growing rate of scientific output.*

Jan Bremer (Medical student at University Medical Center Hamburg-Eppendorf)*: This AI-based Literature Review will lay the foundation for producing more transparent and reproducible results not subject to human biases in the future, which is important across scientific disciplines. In the Covid-19 pandemic, this tool will facilitate the research in the medical community by fast and comprehensive management of the present flood of data.*

Justin Zaremba (Medical student at EVMS)*: Searching for answers to scientific questions can be incredibly time consuming and effort is often wasted analyzing sources that are ultimately irrelevant. Having a tool where I can type in a question and immediately be provided with a list of sources along with the relevant supporting data would not only simplify answering questions that may arise during my education, it would also help me make informed decisions as a future physician*

© Kaggle is bridging the gap between AI and medicine. Learn more about research's role.

https://www.springernature.com/gp/researchers/the-source/blog/blogposts-communicating-research/kaggle-is-bridging-the-gap-between-ai-and-medicine/17995890

Target completion

4 months

Use of funds

$170,000 grant will be allocated throughout 4 months period to cover the costs of cloud computing infrastructure ($10,000); to cover legal and data compliance consulting services ($20,000); to cover basic communication tools including Slack, Zoom, Zapier and others ($10,000); and a compensation to contributors to switch to a full-time commitment required to achieve our ambitious timeframe, including consulting and contracting services ($130,000).

Other funding

$5000 - Google Cloud Platform cloud computing credits

$4000 - Amazon Web Services cloud computing credits

Virtual machines from NASA Jet Propulsion Laboratory (JPL)

Team Highlights

Janosch Ortmann - Assistant Professor in Data Science and Business Intelligence, ESG UQÀM

Janosch's current research includes applications in precision medicine, humanitarian logistics, the design of transport networks and real options. He's currently involved with precision cancer medicine to improve patient outcomes and health spending, he is developing models to understand how a patient's genetic code relates to the success of specific cancer treatments. This work is done in collaboration with the labs of Benjamin Haibe-Kains and Anna Goldenberg, using mouse models and Gaussian Process Regression to predict tumour response to therapy.

- Master in Mathematics (2008) Oxford University (UK)
- PhD in Mathematics (2013) Warwick University (UK)
- Postdoctoral Fellow in Mathematics (2012-2015), University of Toronto
- Postdoctoral Fellow in Mathematics / Mathematical Physics (2015-2017), Université de Montréal
- Postdoctoral Fellow in Mathematical Physics (2017), Concordia University
- Assistant Professor, Data Science and Business Intelligence, Université du Québec à Montréal (UQAM, since 2017)

Dan Sosa - Ph.D. Candidate in Biomedical Informatics at Stanford University

Completed his Bachelor's and M.Eng. at MIT studying Computer Science/Molecular Biology and Management. Currently he is a second-year PhD student in the Biomedical Informatics (BMI) training program. His research interests include generating principled hypotheses about drug repurposing with network-based methods and using chemoinformatics to understand known drug binding interactions and to aid in drug discovery.

Links:

- https://www.linkedin.com/in/dan-sosa-34392064/
- https://www.worldscientific.com/doi/abs/10.1142/9789811215636_0041

Christine Chen, MD MPH - PhD Candidate & Assistant Policy Researcher at RAND Corporation | Data Scientist | Team Lead at CoronaWhy

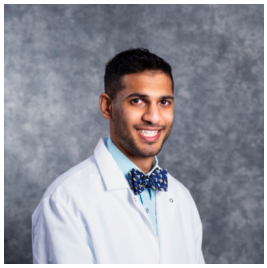
Christine Chen, Research Coordinator (Santa Monica, CA)

- Coming from a multidisciplinary background, her role is to bridge CoronaWhy's technical capabilities and academic research, connecting team members to tasks where they can utilize their talents
- She is a doctoral candidate at the Pardee RAND Graduate School and an assistant policy researcher at RAND.
- Received her M.D. from National Taiwan University and her M.P.H. from the Harvard T.H. Chan School of Public Health.
- Prior to joining Pardee RAND, she was a research analyst at Harvard, where she worked on a project with the Taiwanese National Health Insurance Administration evaluating the performance of Taiwan's health system.
- Conducted systematic reviews on treatments for osteoarthritis, dietary recommendations for sodium and potassium intakes, obesity policies, vaccine beliefs; and a meta-analysis on medication-assisted treatments for opioid addiction

Links:

- https://www.researchgate.net/profile/Christine_Chen26/research?ev=brs_act
- https://www.linkedin.com/in/christine-chen-md-mph-94082370/

Dr Tayab Waseem - Public Policy Fellow for the AAI; Director of Medical Informatics and AI Integration at WMRC



Dr. Waseem has a PhD in Immunology and is a Public Policy Fellow for the American Association of Immunologists. He's the Director of Medical Informatics and AI Integration at the Wagner Macula & Retina Centers and a medical student at Eastern Virginia Medical School. Author of an IRB approved study on the use of AI-powered literature review tools by medical researchers.

Links:

- https://www.linkedin.com/in/tayab-waseem-978449172/

Vyacheslav Tykhonov, Senior Information Scientist at DANS-KNAW | Director of CoronaWhy Data Infrastructure and Services

Vyacheslav Tykhonov is a Senior Information Scientist at the Data Archiving and Networked Services (DANS), an institute of the Royal Netherlands Academy of Arts and Sciences (KNAW) and Dutch funding organization NWO. He is serving as lead developer of the Dataverse SSHOC project, coordinating joint collaboration between European countries for the internalization and further Cloud development of Dataverse, a general-purpose data repository built on open-source software that is intended for sharing and facilitating citation and reuse of research data. Vyacheslav is actively involved in the development of a data platform for the European Open Science Cloud (EOSC), an open, trusted, virtual, federated environment in Europe to store, share and re-use research data across borders.

Vyacheslav is also creating digital tools and new services for the Digital Humanities based on the innovative information technologies. He is actively involved in a number of national and European Union projects and large research infrastructures like Clio-Infra, CLARIAH, EOSC Synergy, PARTHENOS and CESSDA. As a member of DANS-KNAW Research and Innovation group he is working on Linked Data and Semantic Web technologies, including knowledge representation and query languages.

Links:

- https://www.linkedin.com/in/vyacheslavtikhonov

Artur Kiulian - Serial Entrepreneur, Technologist

Artur is a serial entrepreneur, engineer, speaker and author. He built out infrastructure and technology for 70+ companies across his career and has taken 50+ early stage startups to market. Artur has a Masters in Systems of Artificial Intelligence, is a member of Forbes Technology Council and wrote an Amazon best-seller “Robot is The Boss” about practical applications of AI in business.

Links:

- https://www.arturkiulian.com/
- https://www.linkedin.com/in/artur-kiulian/

Advisors

Dr. Byron Wallace (The IRB for the utility study is being run through him at Northeastern University)

Dr. Wallace is an assistant professor at Northeastern University in the Khoury College of Computer Sciences. He is the Director of the BS in Data Science program. He holds an adjunct appointment at Brown University, where he is affiliated with the Center for Evidence Synthesis in Health.

Links:

- https://www.khoury.northeastern.edu/people/byron-wallace/
- https://scholar.google.com/citations?user=KTzRHmwAAAAJ&hl=en

Benjamin M. Gyori - Research Associate in Therapeutic Science at Harvard Medical School

Ben is a Research Associate in Therapeutic Science at the Laboratory of Systems Pharmacology, Harvard Medical School. His research is at the intersection of systems biology, bioinformatics and artificial intelligence, and aims to understand how biological cells work and react to drugs and environmental signals using computational approaches. Ben developed INDRA, a software tool which automatically assembles biochemical mechanisms extracted from the scientific literature into explanatory models. He is also working on a human-machine communication system which allows scientists to interact with a computer partner to construct and test hypotheses about molecular systems. Ben obtained his Ph.D. in computational systems biology from the National University of Singapore.

Links:

- https://www.linkedin.com/in/benjamin-m-gyori-0519b630/
- https://scholar.harvard.edu/bgyori/home
- https://scholar.google.com/citations?user=kH9LOHMAAAAJ&hl=en&oi=ao

Serhii Myroshnychenko - MS, Epidemiology & Biostatistics, Pharma Marketing MBA Fellow

Epidemiologist and an expert in Healthcare & Pharmaceuticals Market Research, skilled in such areas as:

- Public Health Research
- Commercial Pharma Analytics

- Population Health Management
- Biostatistics
- Clinical Research
- Data Science & Advanced Analytics
- Telehealth and mHealth

In his career journey, Serhii has achieved the following results:

- Participated in Telemedicine solution development and launch as a Product Manager;
- Worked in the Infectious Epidemiology projects (HIV and HCV Interventions and Studies)
- Delivered over 100+ Commercial Pharma Analytics Studies (Patient Journey, Drug Adherence, Regimen Studies, Provider Targeting, Provider & Patient Segmentation, Market Sizing Studies)

Links:

- https://www.linkedin.com/in/smyroshnychenko/

Jack Park - AI Researcher and Open-Source Advocate

Jack Park is a computer scientist working in the fields of artificial and collective intelligence. He created, edited, and co-authored the book XML Topic Maps: Creating and Using Topic Maps for the Web, was a Ph.D. student researching the topic of knowledge federation applied to hypermedia discourse, and designs and builds software platforms for knowledge gardening. He was a research scientist at SRI International working on their Cognitive Assistant that Learns and Organizes (CALO) project, and authored and co-authored several conference papers on the subjects of topic mapping and semantic desktop applications for collective intelligence.

Jack now supports organisations and individuals to co-create business models around not-for-profit and b-corporations, and to co-create open source software platforms.

Links:

- https://www.linkedin.com/in/jackpark/
- http://community-intelligence.com/about/team/jack-park/
- http://www.ai.sri.com/software/IRIS

Emad Mostaque - CAIAC & HAI

Leading CAIAC initiative, the partnership between the private sector, academia, government, and multilateral institutions, including UNESCO, the World Bank, the WHO and UN Global Pulse, offering unique

perspectives and solutions on the roadblocks to addressing the COVID-19 pandemic

Links:

- https://hai.stanford.edu/people/emad-mostaque

Jeremy Zucker - Computational Biologist at Pacific Northwest National Laboratory

Jeremy Zucker is a computational biologist at Pacific Northwest National Laboratory. He has over 15 years of experience integrating high-throughput systems biology data with computational methods to study symbiosis, circadian rhythms, evolution, metabolic engineering, human health and infectious disease. Jeremy's current research focuses on applying causal reasoning to predict the determinants of viral pathogenesis and identify targets for medical countermeasures.

Links:

- https://www.linkedin.com/in/jeremyzucker/
- https://scholar.google.com/citations?user=fqTs4kIAAAAJ&hl=en&oi=pll

Charles Tapley Hoyt - Computational Biologist at Enveda Therapeutics

Charles leads the knowledge graph and AI/ML platforms at Enveda Therapeutics to support the identification of novel active phytochemicals from plants and bacteria. His current research is in the applications of knowledge graph embeddings to biological tasks such as drug repositioning, proteochemometrics, toxicology, target prioritization, and precision medicine. He is a member of the Biological Expression Language (BEL) Committee and maintains PyBEL, the software ecosystem supporting the compilation, integration, and analysis of biological knowledge graphs encoded in BEL. He received his Ph.D. in Computational Life Sciences from the University of Bonn.

Links:

- https://cthoyt.com
- https://www.linkedin.com/in/cthoyt/
- https://scholar.google.com/citations?user=PjrpzUIAAAAJ