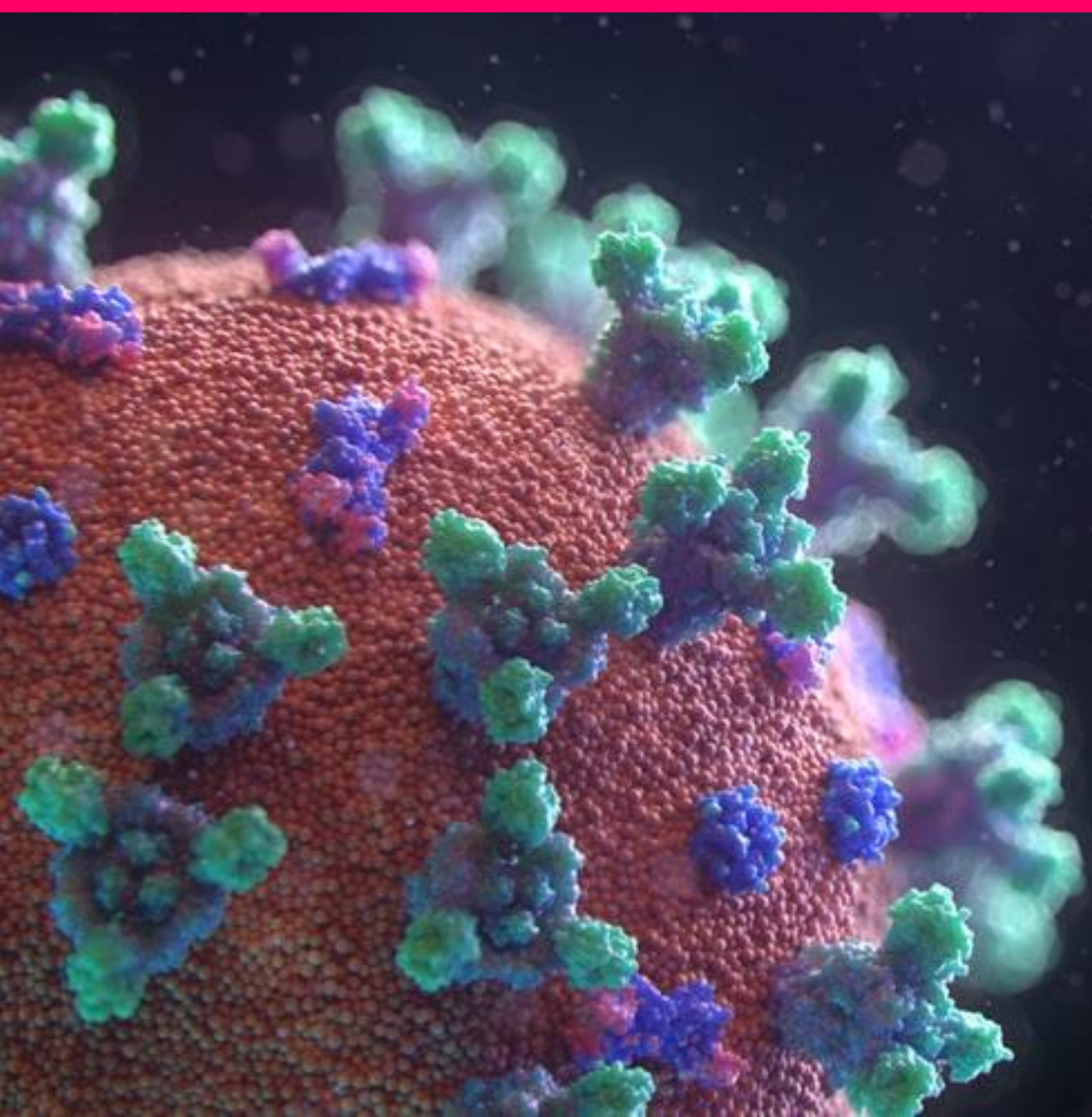

Masters Thesis Overview
COVID-19 Research Cockpit

Jon-Paul Boyd

18th August 2020

DE MONTFORT UNIVERSITY LEICESTER

Introduction About Me & Project

Experienced lead software developer and technical architect with successful track record delivering complex solutions for large enterprise

Successfully completed all 8 Masters Intelligent Systems modules all graded distinction

Passion for A.I. in health sciences with paper classifying breast cancer tumours using fuzzy logic (outperforming neural networks and other ML approaches) being submitted to journal "*Applied Soft Computing*" for publication

Be part of the fight hence I want to use skills and effort for something good, worthy and helpful by choosing a meaningful problem domain

Project status terms of reference complete, knowledge review in progress

Welcome support for scientific advice (understand medical research ways of working, semantics, other information sources), **design validation** (results, UX), **cloud costs** (estimated 5K USD for project lifetime to May 2021, including design, development, deploy, validate and demonstrate)

Background The COVID-19 Pandemic

Causes respiratory infections with common symptoms fever, dry cough and tiredness

Outbreak in Wuhan, China in December 2019 and now a pandemic affecting 188 countries globally (JHU)

774,379 deaths and 21,903,341 confirmed cases globally as reported by JHU (18th August 2020)

Outcomes range from truly asymptomatic to mild, months-long complications and death

High risk individuals include the elderly and those with existing conditions ranging from cancer to obesity to sickle cell disease

Covid-19 is complicated low child deaths, increased blood stickiness, weakened heart muscles, smell/taste loss, autoimmune and mood disorders

Some immunity against reinfection however unknown how long the immune response lasts. Individuals can be infected more than once

Vaccine development is accelerating with compressed and overlapping trials, emergency use licenses

Global Recession third of population into lockdown to slow spread. Economy shrinks 5.2%, deepest recession since WWII (World Bank)

Winter could be worse with second wave expected given more contact in enclosed spaces, overlap with flu

Problems COVID-19 Research

Analysis paralysis due to volume of COVID-19 research literature doubling every 20 days, causing over-analysis, indecision and fear of mistakes

Rush to publication impacts quality with many papers only commentaries, pre-prints not peer reviewed, unoriginal or poor quality

Traditional search engines struggle as research-driven queries need to go beyond surface level keyword matching over documents only, with word meaning and order important

Solution Build A COVID-19 Research Cockpit

Help researchers focus their efforts with a Q&A solution that finds studies most relevant to them, going beyond keyword search only

Connect insights across multiple siloed sources to support the medical community finding best answers to questions about the pandemic

Enable next stage actions and decision-making support by complimenting article retrieval with information extraction

Composite ranking from blending data sources to determine best quality articles

Automated literature summarization generates an article summary relevant to the question, supporting new insight discovery

Topic classification will automate association of multiple topics to an article. A word bubble charts common terms across all result documents sized by frequency

Topic clustering will support intuition and analysis of similarity and distance between topics

Filters for journal, topics, terms, author, affiliation will facilitate iterated refinement of results

Entity graph analysis will show connections between topics, institution affiliation, authors, vaccines etc

Time series visualisation showing number of articles published over time according to selected facets (topics, terms, journal, affiliation, author etc). Indicates emerging trends and directions. Adjusting the timeline (from/to date) updates article inclusion and associated facet lists

Technical Highlights *

CORD-19 dataset of 68K published academic articles on coronaviruses compiled by Google, Allen Institute for AI and other partners

Augment CORD-19 articles with other sources by linking article clinical trial ids to WHO International Clinical Trials Registry Platform (ICTRP), ResearchGate, Google Scholar, social media, news outlets etc. Provides complementary information and supports ranking mechanisms with quality indicators (citations, date published, critique)

Pre-processing remove duplicates, irrelevant documents without key terms, abstracts only, non-English language, stop words. Tokenization (segment into words) and lemmatization (root word). Create metadata for journal, year published, peer review status etc

Leverage scispacy a specialized package to process scientific text, able to normalize technical names including chemical elements and drug names

Keyword search engine allows complex queries with Boolean operators and search on specific fields (title, authors etc). Ranking function such as Okapi BM25 estimates relevance of documents to given query

Unsupervised topic modelling to mine the digital article collection for hidden semantic structure. Learns the abstract topics, where a topic is a distribution over words and each document is a mix of a number of topics. Algorithms might include LDA (Latent Derilicht Analysis) and NMF (Non-negative Matrix Factorization)

Composite ranking further augmented by including additional documents with neighbouring topic distribution to document list generated by keyword search engine

Generic design facilitates exploration of other research problem domains (e.g. upload documents/datasets, alternative topic model plug-ins)

Cloud-hosted application on the Microsoft Azure platform. Persist digital collection, accessibility, scale topic model training, functions as services