**Harvard Data Science Review • Special Issue 1: COVID-19:
Unprecedented Challenges and Chances**

Knowledge Graphs and COVID-19: Opportunities, Challenges, and Implementation

Mayank Kejriwal¹

¹Information Sciences Institute, Daniel J Epstein Department of Industrial and Systems Engineering, Viterbi School of Engineering, University of Southern California

Published on: Dec 01, 2020

DOI: <https://doi.org/10.1162/99608f92.e45650b8>

License: [Creative Commons Attribution 4.0 International License \(CC-BY 4.0\)](#)

ABSTRACT

The COVID-19 pandemic has been truly global and multidimensional in scope, with ramifications extending well beyond health. Yet, unlike previous crises, there is hope that timely release of relevant data sets, as well as advents in AI (artificial intelligence) technology, could lead to compressed timescales in finding a vaccine or cure. Despite the huge existing body of academic literature on the coronavirus family, searching through such a corpus, including new research that has emerged in the wake of the crisis, is a daunting task even for experts. Simple keyword search over such corpora is insufficient for experts who want answers to questions that require linking together multiple pieces of information across documents. In this article, we review an innovative AI technology called a *knowledge graph* (KG) that could be used to fulfill such complex information needs. We detail the potential for KGs to play an important role in the fight against COVID-19. We also cover challenges and ongoing collaborative implementations of COVID-19 KGs in industry and academia.

Keywords: question answering, domain-specific search, information retrieval

1. Introduction

In a recent, instructive article titled “Hoping to Understand the Virus, Everyone Is Parsing a Mountain of Data,” the *New York Times* correspondent Julie Bosman uses a striking example of recent coronavirus case counts to show how difficult it is to draw conclusions without analyzing multiple data points in the right context (Bosman, 2020). Even when the data is not contradictory, the linkages within (and between) multiple data sets and sources imply that most questions cannot be answered simply by ‘skimming’ the data or doing a superficial ‘keyword’ search.

Turning data into knowledge is not a trivial problem, even for humans. For machines, the problem is much harder, especially at scale. Yet, if realized (even partially), the benefits are enormous. Machines with knowledge could augment our own capabilities by serving as assistive subject matter experts (Das et al., 2020), but with the subject matter spanning millions of documents that a single human (or group of humans) would never be able to consume or parse in a lifetime.

More than 70 years ago, Vannevar Bush, then Director of the Office of Scientific Research and Development (a federal agency created to coordinate scientific research for military purposes during World War II, and subsequently discontinued), wrote about the possibility of a “future device...in which an individual stores all his books, records, and communications, and which is mechanized so that it may be consulted with exceeding speed and flexibility. It is an enlarged intimate supplement to his memory” (Bush, 1945). Written in 1945, his vision predates the Internet, artificial intelligence (AI), and even modern computing by many decades. However, despite many inventions, we are still far from realizing the spirit of Bush’s vision, as evidenced by a

recently concluded multimillion-dollar program instituted by the U.S. Department of Defense (Fox-Brewster, 2015).

The problem is that, for all their speed and ‘smarts,’ computers are still far from being a natural and ‘intimate’ supplement to our memory in that they are not able to *understand*, let alone *answer*, the kinds of questions over ‘mountains’ of data that subject matter experts would have been able to answer if only they had been able to read, process, and retain that data. Put another way, machines can *process* great amounts of *data*, but the processing does not necessarily yield *knowledge* that is critical for solving real-world problems (Robbins, 2019).

A real-world example may be a question such as ‘*In what types of cells is SARS-CoV-2 receptor ACE2 primarily expressed?*’ Despite the fact that the answer is available in at least one paper on the Web (Lukassen et al., 2020), and the fact that the search engines have processed and indexed these papers, the best that a search engine can do is to display a list of ‘relevant’ webpages that the user would have to navigate and read to get the answer. Where a question involves multiple facts and inferences, the situation gets progressively worse. For a long time (and to an extent, even today), this resulted in rather primitive search mechanisms, that is, the question above would have to be expressed (or would be interpreted by a search engine) as the keyword query ‘cells SARS-CoV2 ACE2 express.’

Unsurprisingly, keywords are inherently limited in expressing sophisticated user needs: in going from the question to the keywords in the previous, we have essentially purged the question of its ‘semantics.’ Recently, search engines (and researchers in information retrieval) have improved the abilities of their systems to answer questions more directly (Lukovnikov et al., 2017), and extract more precise information from webpages. Many improvements can be traced to sustained investment in the kinds of knowledge-centric AI technologies that we describe in this overview (Lockard et al., 2018; Singhal, 2012). However, many challenges still remain. For instance, Google is still unable to provide a specific answer to the example question we provided.

In fact, building machines that understand and work with knowledge, rather than just data, is one of the holy grails of general AI (Russell & Norvig, 2009). If such programs were to become available, even within limited domains, they would significantly accelerate scientific progress by providing answers to complex questions that may today require many hours of reading, even by subject matter experts. In the specific context of COVID-19 alone, such an assistive technology would prove invaluable to the scientists who are currently working in this area.

In this article, we review a novel kind of AI technology called a *knowledge graph* (KG) that is designed to bring us one step closer to Vannevar Bush’s original vision. We begin with a background and methodological overview of knowledge graphs and their construction (Section 2), followed by milestones in their short but highly significant development (Section 3). In the second half of the article, we discuss the opportunities for

KGs in a COVID-19 world (Section 4), current obstacles and challenges (Section 5), and implementations of efforts already underway (Section 6). Section 7 concludes with some closing thoughts.

2. Knowledge Graphs: Background

We begin our discussion on knowledge graphs with a running example of a small KG fragment (Figure 1) in the *academic* domain comprising scientific papers, authors, and other important details such as the venue or journal where the publication was published. We use actual papers from the literature on the coronavirus family for the illustration.

Figure 1. A knowledge graph (KG) fragment from the citation/academic domain. The details shown here are accurate (we use ... to indicate incompleteness of illustration, e.g., in authorship). The shaded ovals are identifiers of *entities*, while the rectangular boxes are known as *literals* and typically represent attributes of entities such as dates, strings, and numbers. The directed, labeled *edges* represent relations between entities ('cites,' 'published in') or attributes of entities ('author,' 'title').

Figure 1 contains three important types of objects:

- *Entities*, illustrated by shaded oval nodes, express the key objects of interest in the KG. In the running example, these objects are papers and journals.
- *Attributes*, illustrated by rectangles, express the attributes of entities in the domain. Some attributes (such as a paper's authors) could have been modeled as entities, but since they are not the 'primary' objects of interest in this domain, were chosen by the modeler to serve as attributes instead.
- *Relations*, which are the directed, labeled edges (or the 'arrows' in the figure) connecting (i) two entities, or (ii) expressing the fact that an entity has an attribute. Note that attributes may not have 'outgoing' relations, that is, it is invalid to have an arrow protruding from a rectangle in Figure 1.

Not only does the KG contain entities, attributes, and relations, but it also contains *constraints* that are implicit in Figure 1, but can be (and in practice, are) explicitly declared using a ‘domain ontology’ (Missikoff et al., 2002). For example, the commonsense notion that the ‘author’ relation should only apply to papers, and the ‘editor’ relation should only apply to journals can be *enforced* in the KG using formal, declarative rules in the ontology. While a full description of KG formalism is beyond the scope of this article, we use this example to emphasize that KGs are not random combinations of elements, which is an important feature that ‘computational reasoning engines’ often drawn upon when making complex inferences using an initial set of facts (Parsia & Sirin, 2004). Put more broadly, KGs have an inherent notion of *structure*, which allows us to declare complex types of entities and attributes, their interrelationships, and the constraints imposed on them (Kejriwal, 2019).

2.1. KG Structure and Search

The structure in a knowledge graph is vital for machines that, for all the advancements in AI, still do not interpret natural languages such as English in the same way that humans do (Nadkarni et al., 2011). To a machine, a ‘query’ such as ‘How many coronavirus papers have been published in March by Johns Hopkins?’ carries no intrinsic meaning, and is treated as a ‘sequence’ of words or of numerical encodings of words or phrases. Even when these representations can be used to make accurate predictions, they require complex interpretation if they are to be used to generate knowledge understandable to humans. Knowledge graphs have completely overhauled search by instead recognizing that queries, such as in the question above, contain entities (‘Johns Hopkins,’ ‘coronavirus’), concepts (‘papers’), and relations (‘published in March’). Furthermore, the answer is constrained to be a number (a ‘literal’).

By constructing a KG and teaching a search engine how to interface with it, queries and query responses (historically just documents and webpages) can both be decomposed into these richer semantic units, which makes question answering of the kind motivated in the introduction both theoretically possible, and of high-enough quality to be feasible (Lukovnikov et al., 2017).

Unsurprisingly, prior to the advent of KG technology, search engines were mostly capable of handling keyword queries rather than well-formed questions (or question fragments). Search results improved over time for common types of queries, as companies like Google were able to leverage millions of click-logs and other data to train machine learning models to do a better job ranking webpages in response to queries. Nevertheless, the keyword limitation persisted, as at their core, search engines still could not understand the *semantics* of queries. This is why, when the Google Knowledge Graph was first publicized in 2011 (Section 2.3), it promoted the technology with a simple catchphrase ‘Things, not Strings’ (Singhal, 2012). Those three words succinctly capture the essence of why KGs have upended search as we know it.

2.2. Methodologies and Challenges in Constructing Knowledge Graphs

Given raw data (such as a set of scientific articles or even a database), constructing a KG such as in Figure 1 involves a set of methodologies. The main steps are shown in Figure 2. While there is considerable freedom in which methods to adopt for the implementation of these steps (there are hundreds of algorithms available for each box in Figure 2), the *order* of steps shown in the figure is relatively standard (Kejriwal, 2019).

Preliminary steps include acquiring and cleaning the raw data, and designing the domain ontology, which contains the *types* of entities of interest in the domain (e.g., papers, journals), the relationships connecting those entities, attributes associated with these entity types (e.g., ‘date published,’ ‘title’) and constraints. These preliminary steps may be challenging or straightforward, depending on the circumstances. In some cases, the data is already available and does not need to be ‘crawled’ over the Internet or other sources. Nevertheless, data cleaning could be challenging, especially if multiple sources and formats are involved. Similarly, a domain ontology can be challenging to build from scratch if multiple use-cases have to be accommodated. It is sometimes not uncommon for domain ontologies to contain many hundreds, if not thousands, of entity types, and as many constraints.

Figure 2. A simplified workflow illustrating the methodologies for constructing a domain-specific knowledge graph (KG).

Other steps fall within traditional AI research, and most have been researched for many decades. *Information extraction (IE)* involves extracting entities and relations (usually semiautomatically) from the raw data (Nadeau & Sekine, 2007). For example, given a set of biology articles, an ideal IE system would not only extract entities such as proteins, genes, and other entities of interest, but also relations within them. It is important to note that the goal is not to extract ‘all’ entities and relations but only those that we know to be relevant to the domain (specified in the ontology). *Co-reference resolution* is another step that some, but not all, workflows may execute (Raghunathan et al., 2010). Co-reference resolution is the problem of resolving

‘pronouns’ and other such words and phrases to their canonical mentions, in an attempt to avoid needless duplication and obtain higher quality data.

The output of the blue box in Figure 2 should be thought of as the ‘first draft’ of the KG. This first draft has many flaws, the single biggest one of which is a severe duplication problem. For example, the IE may *independently* extract ‘RNA’ and ‘Ribonucleic Acid’ from one or more documents, assigning them different identifiers since they are thought by the system to be different entities. *Entity resolution* is the problem of automatically grouping entities that are believed to be the same underlying entity. It is a hard problem that has emerged in multiple computational communities, including databases, semantic web, and machine learning (Getoor & Machanavajjhala, 2012; Kejriwal & Miranker, 2015). After 50 years of research, recent solutions to the problem have been particularly encouraging, and shown to be useful in real-world architectures.

The KG may also contain noise by way of *incorrectly* extracted relations and entities. In some cases, the correct entity is extracted but its type is wrong (a protein is erroneously declared by the IE to be a gene). So-called ‘knowledge graph identification’ algorithms are necessary for making corrections, as well as for removing erroneous links. These algorithms rely on a range of methodologies (Pujara et al., 2013), but recently, deep-learning techniques such as *knowledge graph embeddings*, have proven to be fast-improving and effective solutions (Wang et al., 2017).

Once the knowledge graph has been constructed and identified (or ‘completed’) in this fashion, it is typically stored in a responsive and specialized *infrastructure* such as a graph database (Angles & Gutierrez, 2008). It must then be exposed to applications such as search engines and question-answering interfaces (such as chatbots). In reviewing important milestones and developments in Section 3, we find that a range of communities have now managed to successfully use knowledge graphs to solve complex problems.

3. Important Developments and Milestones in Knowledge Graph Research

The previous section leads one to believe (and rightly so) that constructing KGs involves work, sometimes considerably so. What makes it worthwhile? While we briefly mentioned some use-cases earlier, this section covers some important developments and milestones in applied KG research.

3.1. Semantic Web Search

Probably the most influential milestone in modern knowledge graph research is the emergence of *semantic search* at an industrial scale. Most individuals do not realize that they have already benefited from KG research if they have used Google in the last 10 years. The *Google Knowledge Graph* (Singhal, 2012) ensures that, when a user inputs a phrasal query such as ‘places to visit in Los Angeles’ in the search engine, she would not merely see a simple list of webpages (ranked by relevance, as determined by Google’s internal methodology), but instead, an actual list of *places* to visit in Los Angeles (Figure 3).

While the most visible (and early) milestone of KG research, the Google Knowledge Graph is preceded by many years of research into KGs in the Semantic Web and broader AI community (Berners-Lee et al., 2001). A foundational goal of the former is to transform the Web by turning it into a Web of interlinked entities (called ‘Web of Linked Data’), rather than a Web of hyperlinked documents. This is best described by an article written by Tim Berners-Lee, the inventor of the World Wide Web, shortly before the emergence of the Google Knowledge Graph (Bizer et al., 2011). Hundreds of papers have since been published on making Linked Data into a practical reality. The synthesis lectures by Heath and Bizer (2011) provide a good overview of the key tenets of the research.

The screenshot shows a Google search interface with the query "places to visit in los angeles". The search results include a "COVID-19 alert" about potential travel impacts, a list of "Top sights in Los Angeles" (Disneyland Park, Universal Studios Hollywood, Hollywood Sign), and a "Knowledge panel" for Los Angeles. The knowledge panel provides detailed information about the city, including its area, population, weather, and a list of "Colleges and Universities".

Search Query: places to visit in los angeles

COVID-19 alert: Potential impact on travel. Information for travelers: www.cdc.gov/coronavirus/2019-ncov/travelers/travel-in-the-us.html

Top sights in Los Angeles:

- Disneyland Park:** 4.6 ★★★★★ (111,017). Famed amusement park with 8 themed lands.
- Universal Studios Hollywood:** 4.6 ★★★★★ (107,851). Big movie-themed amusement park.
- Hollywood Sign:** 4.5 ★★★★★ (6,963). Iconic letters atop Mount Lee since 1923.

Los Angeles Knowledge Panel:

- City in California**
- Description:** Los Angeles is a sprawling Southern California city and the center of the nation's film and television industry. Near its iconic Hollywood sign, studios such as Paramount Pictures, Universal and Warner Brothers offer behind-the-scenes tours. On Hollywood Boulevard, TCL Chinese Theatre displays celebrities' hand- and footprints, the Walk of Fame honors thousands of luminaries and vendors sell maps to stars' homes.
- Area:** 503 mi²
- Metropolitan area:** 4,850 mi²
- Weather:** 85°F (29°C), Wind S at 3 mph (5 km/h), 39% Humidity
- Zip code:** 90019
- Population:** 3.99 million (2018)
- Local time:** Thursday 12:47 PM
- Mayor:** Eric Garcetti
- Plan a trip:** Los Angeles travel guide
- Hotels:** 3-star hotel averaging \$159, 5-star averaging \$522
- Upcoming Events:**
- Colleges and Universities:** View 25+ more

People also ask: Where should I visit in Los Angeles?

Figure 3. An illustration of the indirect manifestation of the Google Knowledge Graph.

When searching for a query such as 'places to visit in Los Angeles,' the search engine produces both a list of such places, a travel alert, the 'traditional' ranked list of webpages (that it would produce prior to the Google Knowledge Graph's invention), as well as a 'knowledge panel' on the side that correctly displays information on the city of Los Angeles.

3.2. Domain-Specific Knowledge Graphs

While efforts like the Google Knowledge Graph are laudable, many organizations or groups specialize in particular domains (e.g., e-commerce, publications) and want to enable rich search applications in those domains. Domain-specific knowledge graphs, constructed using the set of chained methodologies shown in Figure 2, are a means for doing so (Kejriwal, 2019). For example, in the e-commerce domain, enterprises such as Amazon have been building out a *Product Graph* (Krishnan, 2018), a KG-like system for delivering a better user experience in the e-commerce realm (including, among other things, the increased probability that a user

will find and purchase items they are looking for). In another example, Microsoft and LinkedIn (acquired by Microsoft) have also been expanding the LinkedIn graph to offer better utility and recommendations (e.g., for jobs and professional connections) to ordinary users and recruiters alike (He et al., 2020). Many other examples exist and continue to grow (Russell, 2013). In Section 4, we present the notion of a COVID-19 KG as yet another example of a domain-specific KG.

3.3. Knowledge Graphs in Science and Society

Even prior to COVID-19, applications of domain-specific KGs were being explored not only for furthering scientific research but for realizing the vision of ‘AI for social good,’ a movement that has gradually become prominent as the role of AI in society has itself been magnified (Cowls et al., 2019). Domain-specific knowledge graphs have been built in our group at USC, and transitioned to law enforcement for fighting human trafficking (Kejriwal & Szekely, 2017, 2018). In another example, we engaged in a collaborative effort to build KG technology for crisis response under another federally funded effort (Kejriwal & Zhou, 2020). Other knowledge graphs have also been used in specific scientific domains, though sometimes such graphs are designated as ‘ontologies,’ for example, the Gene Ontology (Gene Ontology Consortium, 2015), PubChem (Kim et al., 2016), and geoscience ontologies (Nambiar et al., 2006), to only name a few.

4. Knowledge Graphs and COVID-19: Overview and Opportunities

Given the many milestones in applied KG research, both in for-profit and not-for-profit domains, it is evident that a COVID-19 KG could potentially play an important role for doctors, policymakers, epidemiologists, and other domain experts currently trying to gain deeper insight into the crisis. Table 1 lists some example questions that could be answered by a COVID-19 KG constructed over reasonable data sets. Some of the questions are quite challenging, requiring aggregations and comparisons across spatial regions and time periods. We note that, because of limited manpower and financial resources, the need of the day is to build the KG by using inexpensive, relatively automated solutions, preferably using permissive-license tools from the open-source community.

Table 1. Five example questions that could be serviced by a COVID-19 KG-powered search system but not directly by existing search engines.

Rank the European Union countries in descending order of GDP decline in the first quarter compared to the same quarter in 2019.
In what types of cells is SARS-CoV-2 receptor ACE2 primarily expressed?
In which US city with population greater than 70,000 has there been maximum per capita COVID-19 case count increase in the last 24 hours?

Rank the counties in Southern California in descending order by number of COVID-19 case counts in the last seven days.
--

What is the latest US county to have issued a stay-at-home order?

A COVID-19 KG is another example of a *domain-specific* knowledge graph that was described in Section 3.2 (Kejriwal, 2019). A distinguishing factor in building a COVID-19 KG is that the timescale is far more compressed compared to building an e-commerce KG (for example), which has developed over years of intense research using deep corporate funding (Li et al., 2017). Despite such challenges, realizing a compressed timescale for building a COVID-19 KG has become feasible in no small part due to the lessons learned from those prior efforts, and the increased advent of open-source software and data release (Wright et al., 2020). Next, we take a closer look at opportunities, use-cases, challenges, and ongoing implementations of candidate COVID-19 KGs.

4.1. Inputs to a COVID-19 KG

In describing the workflow and methodology for constructing a domain-specific KG (Figure 2), we stated that the process must begin with a domain ontology and raw data sets over which to construct the KG. While the workflow is not an all-or-nothing proposition and involves considerable flexibility in implementation, it is nonetheless a good starting point. Certainly, it is undisputed that building a KG requires good input data sets. The domain ontology, on the other hand, is less of a concern when dealing with short timespans or limited domains. They also tend to be specific to the use-case, for example, medical researchers looking for a vaccine will benefit from more scientific domain ontologies (the Gene Ontology is a good example), while economists and policymakers will clearly be interested in a separate set of concepts (e.g., geospatial attributes, economic markers such as GDP and unemployment, etc.).

Luckily, due to mobilization of governments and various groups, both private and public, there is no shortage of detailed data sets that could serve as inputs to the KG. In Table 2, we provide examples of representative data sets that, in some cases, are already being used to build a COVID-19 KG.

Table 2. A nonexhaustive list of promising data sets that could be (and in some cases such as in Section 6, already are being) used to build one or more COVID-19 knowledge graphs.

Data Set	Description
----------	-------------

COVID-19 Open Research Dataset (CORD-19)	Scholarly literature about COVID-19, SARS-CoV-2, and the Coronavirus group released by researchers and leaders from the Allen Institute for AI, Chan Zuckerberg Initiative, Georgetown University’s Center for Security and Emerging Technology, Microsoft, and the National Library of Medicine at the National Institutes of Health (Allen Institute for AI, 2020).
Yahoo COVID-19	We describe the implementation of a COVID-19 KG recently announced and released by the Yahoo Knowledge Graph team in a subsequent section. The GitHub repository available on the project page (Nagpal, 2020) contains links to a full list of country-specific and (for the United States) county-specific data sets.
ECDC-COVID-19	The European Centre for Disease Prevention and Control’s (ECDC) Intelligence team has been collecting the number of COVID-19 cases and deaths on a daily basis by collating reports from health authorities worldwide. This is a high-quality source since, each day, a team of epidemiologists screens up to 500 relevant sources to collect the latest figures (EU Open Data Portal, 2020).
Kaggle	Kaggle is an important source for many data sets (including CORD-19) that have been released since the pandemic broke out. It is also an important repository for tools and coding notebooks (Kaggle, 2020).
Lens COVID-19 data set	Free and open data sets of patent documents, scholarly research works metadata, and biological sequences from patents assembled and made available by the Lens in a machine-readable format that is also amenable to intuitive exploration (The Lens, 2020).
2019-nCoV	This data repository is operated by the Johns Hopkins University Center for Systems Science and Engineering (JHU CSSE) and is hosted on GitHub (JHU CSSE, 2020). It is a secondary data set that collects together many data sets from governments and commissions (such as the National Health Commission of the People’s Republic of China, ECDC, etc.).

GeoCoV19	A multilingual Twitter data set released in late May 2020 and comprising more than half a billion multilingual tweets posted over a period of 90 days since February 1, 2020. The authors of the paper (currently available as the arXiv preprint) describing the data set applied a gazetteer-based approach to infer geolocation of tweets (Qazi et al., 2020).
----------	---

Researchers have already been processing some of these data sets using Natural Language Processing (NLP), particularly CORD-19. Recently, for example, Lu Wang et al. (2020) processed the CORD-19 corpus using information extraction algorithms, producing a secondary data set called CORD-NER (Named Entity Recognition). This data set can now be directly used by others, enabling them to bypass the information extraction step in Figure 2, and making it much more efficient to stand up a KG infrastructure. Impressively, CORD-NER covers 75 fine-grained entity types, including common biomedical classes (genes, chemicals, and diseases) and entity types expected to be useful for COVID-19 research studies, including coronaviruses, substrates, immune responses, and viral proteins.

4.2. Greenfield Opportunities

The availability of data, and execution of information extraction algorithms, is only the beginning of the process of constructing a full-fledged COVID-19 KG. To our knowledge, several ‘greenfield’ opportunities still remain unexploited, despite their imminent feasibility given the current state of knowledge graph technology. Following, we describe two promising opportunities.

COVID-19 Social Media KG. An important, perhaps obvious, opportunity, is to make greater use of some of the data sets listed in Table 2. For example, we are still not aware of a concerted infrastructure, either in academia or industry, that has leveraged the massive amounts of *social media* data that have been released so far. While there have been some sporadic attempts to detect COVID-19 misinformation and conspiracy theories on Twitter and other social media platforms, there hasn’t been a focused and large-scale attempt to organize the useful ‘grassroots’ knowledge in social media into a KG (Ahmed et al., 2020; Kouzy et al., 2020).

One reason may be that it is much harder to guarantee quality of a KG that uses social media as a primary data source, especially considering the prevalence of bots and other nefarious sources (Ferrara, 2020). More broadly, a comprehensive integration of data sources that includes economic data, social media, data released by governments, states and municipal counties, and generic databases such as OpenStreetMap and GeoNames (Bennett, 2010; Hahmann & Burghardt, 2010), continues to be the holy grail of any COVID-19 KG effort. This is a true greenfield opportunity that AI researchers should potentially collaborate on.

COVID-19 Metadata KG. A related, and unexploited, greenfield opportunity is a *metadata KG* that contains details, including timestamp, provenance, citation and usage, on software and data resources that could be used by an enterprising engineer to stand up an application for achieving a specific goal. For example, the goal may

be to estimate or model outbreaks of the virus by combining data sets across sources. By using the metadata KG, the engineer or data scientist would not only be able to find relevant sources of high quality but would be able to find software and resources to parse and process those sources, potentially avoiding significant duplication of work that has already been done elsewhere.

A metadata COVID-19 KG is inherently scalable since it only contains *pointers* to the data, rather than a copy of the data itself. It could also serve as a ‘live’ record of all data sets, software, visualizations, and systems that are being continuously released on the Web by many different groups in support of COVID-19 research. Currently, all of these resources are scattered, making it that much more difficult to make effective use of them jointly. In our view, a good metadata KG should not be limited to describing data sources, but should describe resources more broadly, including (as suggested above) software and visualizations.

In our own group at USC (and in collaboration with researchers from across USC), we are designing and pursuing just such a prototype, informally titled as PLETHORA, to support computational systems scientists in *spatiotemporal* data integration. To support spatiotemporal visualizations and reasoning, the data sets in Table 2 need to be ‘linked’ to other data sets that provide detailed location information, including maps. Table 3 presents examples of some data sets that are being considered in PLETHORA’s design. We hope to prototype PLETHORA in the Los Angeles metropolitan area in 2021, and to use it as a model for other cities and regions over a multiyear period.

Table 3. Examples of important resources that will be incorporated into PLETHORA to support a COVID-19 KG infrastructure for spatiotemporal question answering and visualization.

Resource Name	Description
OpenStreetMap	The crowd-sourced, open version of Google Maps
Twitter API	Application Programming Interface for accessing and downloading streaming Twitter data
GeoNames	Point data sets containing millions of geographical names and feature types
Crime Data Explore	Point data sets containing crime types and locations in the United States
Air Data	Point data sets containing outdoor air quality from the U.S. Environmental Protection Agency monitoring stations

Gallup World Poll	A proprietary data set containing global polling data on a variety of issues (especially, well-being, diversity, and inclusion) that can be used to address both longitudinal and cross-sectional research questions
-------------------	--

5. Challenges in COVID-19 KG Construction

For greenfield opportunities to be realized, several important challenges need to be addressed. In carefully surveying the state-of-the-art technology that currently exists for building and using domain-specific KGs, we group these challenges under three broad umbrellas.

5.1. Quality and Scale of KG Construction

The most immediate challenge is the quality of algorithmic outputs produced by the different AI modules in Figure 2. If a COVID-19 KG is not deemed to be of sufficient quality, it will not be used since stakeholders (both scientists and policymakers) would want to be sure that resources are not being wasted by relying on untrustworthy data or algorithms. It is unlikely that any algorithm in knowledge graph construction will reach human-level accuracy in the near future. Therefore, a prudent area of research is in determining just what it means for the quality of an algorithm to be *good enough* to ensure that the KG can be used reliably for the fundamental purpose of semantic search, which has always been its primary utility. Recently, *explainable AI* (especially in the medical domain) has emerged as an important research area in this direction (Holzinger et al., 2017).

In particular, despite many advances, the information extraction (IE) problem that has been consistently mentioned in the article as an important methodological step in KG construction still lags human performance quite significantly. This is especially the case when extracting relations (the labeled arrows in Figure 1) and also events, although enormous progress has been made on both (Smirnova & Cudré-Mauroux, 2018). Furthermore, the quality starts declining rapidly as the domain becomes more ‘unusual,’ and without dispute, COVID-19 would qualify as an unusual domain. While AI researchers have discovered mechanisms through which to further ‘boost’ an algorithm’s quality, for example, by using an ensemble of algorithms, and calibrating the algorithm’s parameters to provide better uncertainty estimates, much more research is necessary.

Another issue that deserves attention is *scale*. Scale is currently not likely to be a severe challenge for building a COVID-19 KG, since the data sets in Table 2 do not fall under a ‘Big Data’ definition (even when combined together). However, we expect that scale will become an issue in later phases of the crisis, as more data is collected and collated across agencies, countries, and organizations. Finally, privacy concerns are also expected to arise in this context, especially with increasing public support of contact tracing and apps released by tech giants such as Apple and Google (Cho et al., 2020).

5.2. Development of Intuitive Frontend Tools and Visualizations

Even with imperfect technology, users can be empowered to have more faith in an algorithm by having access to appropriate frontend tools and visualizations. In fact, an important challenge when deploying domain-specific KGs is *usability* of the final infrastructure. This is less of a challenge when the primary use of the KG is at the *backend* of a ‘larger’ infrastructure as has historically been the case with search engines and e-commerce platforms that incorporate KGs as one of several technologies for enhancing the search and user experience. For those applications, the search interface always has a human-in-the-loop element, and search engines are constantly optimizing by leveraging click logs (of many millions of daily users) and other data collected on the backend. Users and customers are not allowed to *directly* query the KG or measure its quality. For example, no one outside of Google (to our knowledge) has direct access to the Google Knowledge Graph. Similarly, neither customers nor third-party vendors have direct access to Amazon’s Product Graph. Rather, these graphs are exposed on a case-by-case basis when the user conducts an actual search.

In contrast, a COVID-19 KG must directly support the complex queries of domain experts by allowing intuitive inputs (preferably in natural language) and easy ways to visualize and explore outputs. Unfortunately, visualization research in the KG community has not kept pace with research in KG construction.

Encouragingly, the last year has seen a spurt of research in this area, and some of the tools seem ready for broad usage. For example, last year we published and demonstrated a tool called SAVIZ that enables a user to visualize the outputs of machine learning classifiers applied to a social media knowledge graph. We used a KG constructed from Twitter in the aftermath of a crisis, such as the 2015 Gorkha earthquake in Nepal, as a case-study illustrating the applicability of that work (Kejriwal & Zhou, 2019). Other such tools have been released by other groups (He et al., 2019). These tools could potentially be retrofitted to service a broader COVID-19 KG effort by supporting interactive visualization and exploration.

5.3. Social Challenges

There are also social challenges in using any advanced technology, especially when promoting them to stakeholders who are conservative about putting too much faith in the ability of machines to influence society for the better (Vayena et al., 2018). While statistical arguments (and also explainable AI) could potentially be used to alleviate the concerns of rational scientists, epidemiologists, or doctors, it may be far more difficult to assuage or even anticipate general societal fears, such as loss of privacy and workforce due to automation. An element of outreach and education is, therefore, crucial to broader uptake of KGs, and AI in general. This is especially true for communities and developing countries that have been historically underserved (or even counterserved) by technological advances.

In our own experience, we have found that presenting KG systems as *augmented AIs* and emphasizing the human-in-the-loop element built into such systems has allowed us to make significant inroads in several nontechnological domains, most important of which has been the use of KGs to fight human trafficking

(Kejriwal & Szekely, 2018; Kejriwal et al., 2018). The burgeoning movement of AI for social good also provides valuable guidance in this area (Green, 2019). In the long term, this kind of outreach is critical for the technology to counter social challenges with rational discourse.

6. Current Large-Scale Efforts for Constructing a COVID-19 KG

The challenges in Section 5 have not stopped some ambitious groups from launching a COVID-19 knowledge graph into the public domain. Most likely there are other efforts that are underway but are currently private, such as in our own group. We describe three public efforts of which we are aware, and which have a dedicated website for the project.

6.1. Yahoo Knowledge Graph for COVID-19

A notable effort in industry is by the Yahoo Knowledge (YK) team at Verizon Media (which acquired Yahoo a few years ago). The team already had a prominent presence in knowledge graph research even before the acquisition, and has continued to innovate since. When the COVID-19 pandemic started snowballing, the team applied their research in web-scale extraction technologies to start building a COVID-19 KG. They extracted statistics from hundreds of sources around the world into a data set dubbed the YK-COVID-19 data set (Nagpal, 2020). The YK-COVID-19 data set is updated multiple times a day at the time of writing, and provides reports at country, state, and even county levels (conditioned, of course, on availability of data).

The YK-COVID-19 data set has been made available under a Creative Commons CC-BY-NC 4.0 license. Considering this implementation in the context of the challenges we covered earlier, we note that one way that the Yahoo team has attempted to inculcate trust in their system is by claiming to provide website-level provenance for every single statistic in their data set. Furthermore, they have built and released dashboards and APIs to analyze, use, and visualize the data.

6.2. The COVID*GRAPH project

The COVID*GRAPH project, hosted by ODBMS.org, is an interdisciplinary project that emerged very quickly after the pandemic appeared as a global crisis with lockdowns happening in swift succession across nations. We cite this example as a model case study of collaboration between academia and industry in building such KGs, with involvement from researchers and data scientists in the German Center for Diabetes Research, Aarhus University, Kaiser & Preusse, yWorks, Neo4j, and several other organizations (COVID*GRAPH, 2020).

Data sets that are integrated into the graph include many of the data sets listed in Table 1, including CORD-19, 2019-nCoV, and the Lens COVID-19 data set. It also integrates open-source knowledge bases such as the Gene Ontology and the NCBI Gene Database. Therefore, it provides a sophisticated degree of entity resolution for biological entities, allowing for a more complete and densely connected KG to be used by stakeholders.

An important benefit provided by the COVID*GRAPH project is that the knowledge graph is implemented in *Neo4j*. *Neo4j* is an open-source, NoSQL, native-graph database that has been in development since 2003, but was only made publicly available starting 2007 (Webber, 2012). Just like other similar software in this space (motivated by for-profit enterprise needs, but without sacrificing the benefits of open-source community-driven development), *Neo4j* has both a Community Edition and Enterprise Edition of the database. *Neo4j* can be queried using a declarative query language called *Cypher* that is SQL-like but is optimized for graphs and intuitive to use. *Neo4j* can also be used to provide network analytics and visualizations. By using *Neo4j* for modeling, storing, and exposing the KG, the COVID*GRAPH project considerably simplifies adoption by a large body of data scientists and app developers (not to mention, researchers from the Semantic Web and AI communities involved in graph research).

6.3. The COVID-19 Pathophysiology Knowledge Graph

Although described as the ‘COVID-19 Knowledge Graph’ in the preprint recently published describing this KG (Domingo-Fernández et al., 2020), the KG is mainly designed to formalize knowledge derived from publications on the pathophysiology of the COVID-19 virus into a ‘computable, multi-modal, cause-and-effect knowledge model.’ Similar to the two previous case study implementations, the authors used scientific literature that was from open-access and freely available journals. The authors further filtered these papers based on information pertinent to potential drug targets for COVID-19, pathways in which the virus interfered to replicate in the host, and data on viral proteins and functions. The authors also prioritized the articles in terms of the level of information that could be captured using the modeling language that they chose to use for building the KG. Ultimately, the KG presented in the paper contains mechanistic data on COVID-19 published in 145 research articles. It is quite modest by the standards of the KGs in Sections 6.2 and 6.1, containing about 3,954 nodes, 10 entity types, and almost 9,500 relations. However, the KG seems to be very well-maintained and may prove to be a particularly high-quality, curated resources for doctors and medical researchers.

The authors have released this KG under the CC-0 license, and have set up a dedicated GitHub repository and a webpage, both of which may be found in the paper abstract (Domingo-Fernández et al., 2020).

7. Closing Notes

This article aimed to provide a broad and accessible introduction to knowledge graphs, and their potential for equipping domain experts and scientists with crucial insights from the vast (and growing) swathes of data made publicly available since the global onset of the COVID-19 pandemic. This potential is not just theoretical, since at least two groups have already released implementations of COVID-19 KGs that could be used in complementary ways. We believe that this is just the tip of the iceberg, and much more work will be forthcoming over the remainder of the year. In our own group at the University of Southern California, for example, we are ingesting and processing several of the data sets listed in Table 2 into a unified knowledge graph that can be queried and analyzed using intuitive interfaces. The work is funded by federal agencies,

including the Department of Defense. It falls under the same ‘AI for social good’ umbrella as other projects where we have built KGs to facilitate such use-cases as fighting human trafficking and natural disaster response by mining social media data, often in non-English settings.

Most likely there are other academic groups engaged in similar endeavors at the moment. However, the COVID*GRAPH, COVID-19 Pathophysiology Knowledge Graph, and Yahoo COVID-19 KG efforts already illustrate the promise and growth of KGs, since, until as recently as just a couple of years ago, standing up a full-fledged domain-specific KG implementation and public-facing architecture within months of a pandemic would have been considered infeasible. Although there are still many challenges and opportunities to be tackled, the technology has clearly come a long way in just under a decade.

Disclosure Statement

Mayank Kejriwal has no financial or non-financial disclosures to share for this article.

References

Ahmed, W., Vidal-Alaball, J., Downing, J., & Seguí, F. L. (2020). COVID-19 and the 5G conspiracy theory: Social network analysis of Twitter data. *Journal of Medical Internet Research*, 22(5), Article e19458.

<https://doi.org/10.2196/19458>

Allen Institute for AI. (2020). *COVID-19 open research dataset challenge*. Retrieved May 29, 2020, from <https://www.kaggle.com/allen-institute-for-ai/CORD-19-research-challenge>

Angles, R., & Gutierrez, C. (2008). Survey of graph database models. *ACM Computing Surveys*, 40(1), 1–39. <https://doi.org/10.1145/1322432.1322433>

Bennett, J. (2010). *OpenStreetMap*. Packt Publishing Ltd.

Berners-Lee, T., Hendler, J., & Lassila, O. (2001). The semantic web. *Scientific American*, 284(5), 34–43. <https://www.scientificamerican.com/article/the-semantic-web/>

Bizer, C., Heath, T., & Berners-Lee, T. (2011). Linked data: The story so far. In *Semantic services, interoperability and web applications: Emerging concepts* (pp. 205–227). IGI

Global. <https://doi.org/10.4018/978-1-60960-593-3.ch008>

Bosman, J. (2020, July 27). Hoping to understand the virus, everyone is parsing a mountain of data. *The New York Times*. <https://www.nytimes.com/2020/07/27/us/coronavirus-data.html>

Bush, V. (1945). As we may think. *The Atlantic Monthly*, 176(1), 101–108. <https://doi.org/10.3998/3336451.0001.101>

Cho, H., Ippolito, D., & Yu, Y. W. (2020). Contact tracing mobile apps for COVID-19: Privacy considerations and related trade-offs. *arXiv*. <https://doi.org/10.48550/arXiv.2003.11511>

COVID*GRAPH. (2020, March 30). *We build a knowledge graph on COVID-19*. ODBMS. <http://www.odbms.org/2020/03/we-build-a-knowledge-graph-on-covid-19/>

Cowls, J., King, T., Taddeo, M., & Floridi, L. (2019). Designing AI for social good: Seven essential factors. *SSRN*. <https://doi.org/10.2139/ssrn.3388669>

Das, A., Datta, S., Gkioxari, G., Lee, S., Parikh, D., & Batra, D. (2018). Embodied question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops* (pp. 2054–2063). IEEE. <https://doi.org/10.1109/cvpr.2018.00008>

Domingo-Fernandez, D., Baksi, S., Schultz, B., Gadiya, Y., Karki, R., Raschka, T., Eberling, C., Hofmann-Apitius, M., & Kodamullil, A. T. (2020). COVID-19 Knowledge Graph: A computable, multi-modal, cause-and-effect knowledge model of COVID-19 pathophysiology. *BioRxiv*. <https://doi.org/10.1101/2020.04.14.040667>

EU Open Data Portal. (2020). *COVID-19 coronavirus data*. Retrieved May 29, 2020, <https://data.europa.eu/euodp/en/data/dataset/covid-19-coronavirus-data>

Ferrara, E. (2020). What types of COVID-19 conspiracies are populated by Twitter bots? *First Monday*, 25(6), <https://doi.org/10.5210/fm.v25i6.10633>

Fox-Brewster, T. (2015, April 10). Memex in action: Watch DARPA artificial intelligence search for crime on the “Dark Web.” *Forbes*. <https://www.forbes.com/sites/thomasbrewster/2015/04/10/darpa-memex-search-going-open-source-check-it-out/#10bed18d2812>

Gene Ontology Consortium. (2015). Gene Ontology Consortium: Going forward. *Nucleic Acids Research*, 43(D1), D1049–D1056. <https://doi.org/10.1093/nar/gku1179>

Getoor, L., & Machanavajjhala, A. (2012). Entity resolution: Theory, practice & open challenges. *Proceedings of the VLDB Endowment*, 5(12), 2018–2019. <https://doi.org/10.14778/2367502.2367564>

Green, B. (2019). “Good” isn’t good enough. In *Proceedings of the 2019 AI for Social Good Workshop at NeurIPS*. GitHub. <https://aiforsocialgood.github.io/neurips2019/>

Hahmann, S., & Burghardt, D. (2010). Connecting linkedgeodata and geonames in the spatial semantic web. In *6th International GIScience Conference*. Springer.

He, Q., Yang, J., & Shi, B. (2020, April). constructing knowledge graph for social networks in a deep and holistic way. In *Companion Proceedings of the Web Conference 2020* (pp. 307–308). Association for

Computing Machinery. <https://doi.org/10.1145/3366424.3383112>

He, X., Zhang, R., Rizvi, R., Vasilakes, J., Yang, X., Guo, Y., He, Z., Prosperi, M., Huo, J., Alpert, J., & Bian, J. (2019). ALOHA: Developing an interactive graph-based visualization for dietary supplement knowledge graph through user-centered design. *BMC medical informatics and decision making*, 19(4), Article 150. <https://doi.org/10.1186/s12911-019-0857-1>

Heath, T., & Bizer, C. (2011). Linked data: Evolving the web into a global data space. *Synthesis Lectures on the Semantic Web: Theory and Technology*, 1(1), 1–136. <https://doi.org/10.2200/s00334ed1v01y201102wbe001>

Holzinger, A., Biemann, C., Pattichis, C. S., & Kell, D. B. (2017). What do we need to build explainable AI systems for the medical domain? *arXiv*. <https://doi.org/10.48550/arXiv.1712.09923>

Johns Hopkins University Center for Systems Science and Engineering. (2020). *Novel coronavirus (COVID-19) cases, provided by JHU CSSE*. GitHub. Retrieved May 29, 2020, from <https://github.com/CSSEGISandData/COVID-19>

Kaggle. (2020). *Help us better understand COVID-19*. Retrieved May 29, 2020, from <https://www.kaggle.com/covid19>

Kejriwal, M. (2019). *Domain-specific knowledge graph construction*. Springer. <https://doi.org/10.1007/978-3-030-12375-8>

Kejriwal, M., & Miranker, D.P. (2015). An unsupervised instance matcher for schema-free RDF data. *Journal of Web Semantics*, 35(Part 2), 102–123. <https://doi.org/10.1016/j.websem.2015.07.002>

Kejriwal, M., & Szekely, P. (2017). Knowledge graphs for social good: An entity-centric search engine for the human trafficking domain. *IEEE Transactions on Big Data*. <https://doi.org/10.1109/tbdata.2017.2763164>

Kejriwal, M., & Szekely, P. (2018). Technology-assisted investigative search: A case study from an illicit domain. In *Extended Abstracts of the 2018 CHI Conference on Human Factors in Computing Systems* (Paper No. CS17). Association for Computing Machinery. <https://doi.org/10.1145/3170427.3174364>

Kejriwal, M., Szekely, P., & Knoblock, C. (2018). Investigative knowledge discovery for combating illicit activities. *IEEE Intelligent Systems*, 33(1), 53–63. <https://doi.org/10.1109/mis.2018.111144556>

Kejriwal, M., & Zhou, P. (2019). SAVIZ: Interactive exploration and visualization of situation labeling classifiers over crisis social media data. In *Proceedings of the 2019 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining* (pp. 705–708). Association for Computing Machinery. <https://doi.org/10.1145/3341161.3343703>

Kejriwal, M., & Zhou, P. (2020). On detecting urgency in short crisis messages using minimal supervision and transfer learning. *Social Network Analysis and Mining*, 10(1), 1–12. <https://doi.org/10.1007/s13278-020-00670-7>

Kim, S., Thiessen, P. A., Bolton, E. E., Chen, J., Fu, G., Gindulyte, A., Han, L., He J., He, S., Shoemaker, B. A., Wang, J., Yu, B., Zhang, J., & Bryant, S. H. (2016). PubChem substance and compound databases. *Nucleic Acids Research*, 44(D1), D1202–D1213. <https://doi.org/10.1093/nar/gkv951>

Kouzy, R., Abi Jaoude, J., Kraitam, A., El Alam, M. B., Karam, B., Adib, E., Zarka, J., Traboulsi, C., Akl, E. W., & Baddour, K. (2020). Coronavirus goes viral: quantifying the COVID-19 misinformation epidemic on Twitter. *Cureus*, 12(3), Article e7255. <https://doi.org/10.7759/cureus.7255>

Krishnan, A. (2018, August 17). Making search easier: How Amazon’s Product Graph is helping customers find products more easily [Blog post]. *Amazon*. <https://blog.aboutamazon.com/innovation/making-search-easier>

The Lens. (2020). *Human Coronaviruses data initiative*. Retrieved May 29, 2020, from <https://about.lens.org/covid-19/>

Li, F. L., Qiu, M., Chen, H., Wang, X., Gao, X., Huang, J., Ren, J., Zhao, Z., Zhao, W., Wang, L., Jin, G., & Chu, W. (2017). Alime assist: An intelligent assistant for creating an innovative e-commerce experience. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management* (pp. 2495–2498). Association for Computing Machinery. <https://doi.org/10.1145/3132847.3133169>

Lockard, C., Dong, X. L., Einolghozati, A., & Shiralkar, P. (2018). CERES: Distantly supervised relation extraction from the semi-structured web. *Proceedings of the VLDB Endowment*, 11(10), 1084–1096. <https://doi.org/10.14778/3231751.3231758>

Lu Wang, L., Lo, K., Chandrasekhar, Y., Reas, R., Yang, J., Eide, D., Funk, K., Kinney, R., Liu, Z., Merrill, W., Mooney, P., Murdick, D., Rishi, D., Sheehan, J., Shen, Z., Stilson, B., Wade, A. D., Wang, K., Wilhelm, C., Xie, B., ... Kohlmeier, S. (2020). CORD-19: The Covid-19 open research dataset. *arXiv*. <https://doi.org/10.48550/arXiv.2004.10706>

Lukassen, S., Chua, R. L., Trefzer, T., Kahn, N. C., Schneider, M. A., Muley, T., Winter, H., Meister, M., Veith, C., Boots, A. W., Hennig, B. P., Kreuter, M., Conrad, C., & Eils, R. (2020). SARS-CoV-2 receptor ACE 2 and TMPRSS 2 are primarily expressed in bronchial transient secretory cells. *The EMBO Journal*, 39(10), Article e105114. <https://doi.org/10.15252/embj.20105114>

Lukovnikov, D., Fischer, A., Lehmann, J., & Auer, S. (2017). Neural network-based question answering over knowledge graphs on word and character level. In *Proceedings of the 26th International Conference on World Wide Web* (pp. 1211–1220). Association for Computing Machinery. <https://doi.org/10.1145/3038912.3052675>

- Missikoff, M., Navigli, R., & Velardi, P. (2002). The usable ontology: An environment for building and assessing a domain ontology. In I. Horrocks, & J. Hendler (Eds.), *Lecture Notes in Computer Science: Vol. 2342. The Semantic Web—ISWC 2002* (pp. 39–53). Springer. https://doi.org/10.1007/3-540-48005-6_6
- Nadeau, D., & Sekine, S. (2007). A survey of named entity recognition and classification. *Lingvisticae Investigationes*, 30(1), 3–26. <https://doi.org/10.1075/li.30.1.03nad>
- Nadkarni, P. M., Ohno-Machado, L., & Chapman, W. W. (2011). Natural language processing: An introduction. *Journal of the American Medical Informatics Association*, 18(5), 544–551. <https://doi.org/10.1136/amiajnl-2011-000464>
- Nagpal, A. (2020, April 27). Yahoo Knowledge Graph Announces COVID-19 Dataset, API, and Dashboard with Source Attribution [Blog post]. *Yahoo*. <https://developer.yahoo.com/blogs/616566076523839488/>
- Nambiar, U., Ludaescher, B., Lin, K., & Baru, C. (2006). The GEON portal: Accelerating knowledge discovery in the geosciences. In *Proceedings of the 8th Annual ACM International Workshop on Web Information and Data Management* (pp. 83–90). Association for Computing Machinery. <https://doi.org/10.1145/1183550.1183567>
- Parsia, B., & Sirin, E. (2004, November). Pellet: An owl dl reasoner. In *Third International Semantic Web Conference-Poster* (Vol. 18, p. 13). <http://iswc2004.semanticweb.org/posters/PID-ZWSCSLQK-1090286232.pdf>
- Pujara, J., Miao, H., Getoor, L., & Cohen, W. (2013). Knowledge graph identification. In H. Alani, L. Kagal, A. Fokoue, P. Groth, C. Biemann, J. X. Parreira, L. Aroyo, N. Noy, C. Welty, & K. Janowicz (Eds.), *Lecture Notes in Computer Science: Vol. 8218. The Semantic Web – ISWC 2013* (pp. 542–557). Springer. https://doi.org/10.1007/978-3-642-41335-3_34
- Qazi, U., Imran, M., & Ofli, F. (2020). GeoCoV19: A dataset of hundreds of millions of multilingual COVID-19 tweets with location information. *SIGSPATIAL Special*, 12(1), 6–15. <https://doi.org/10.1145/3404111.3404114>
- Raghunathan, K., Lee, H., Rangarajan, S., Chambers, N., Surdeanu, M., Jurafsky, D., & Manning, C. (2010). A multi-pass sieve for coreference resolution. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing* (pp. 492–501). Association for Computational Linguistics. <https://www.aclweb.org/anthology/D10-1048>
- Robbins, S. (2019). AI and the path to envelopment: Knowledge as a first step towards the responsible regulation and use of AI-powered machines. *AI & SOCIETY*, 35(2), 391–400. <https://doi.org/10.1007/s00146-019-00891-1>

Russell, M. A. (2013). *Mining the social web: Data mining Facebook, Twitter, LinkedIn, Google+, GitHub, and more*. O'Reilly Media, Inc.

Russell, S., & Norvig, P. (2009). *Artificial intelligence: A modern approach*. Cambridge University Press.
<https://doi.org/10.1017/S0269888900007724>

Singhal, A. (2012, May 16). Introducing the knowledge graph: Things, not strings [Blog post]. *Google*.
<https://www.blog.google/products/search/introducing-knowledge-graph-things-not/>

Smirnova, A., & Cudré-Mauroux, P. (2018). Relation extraction using distant supervision: A survey. *ACM Computing Surveys*, 51(5), Article 106. <https://doi.org/10.1145/3241741>

Vayena, E., Blasimme, A., & Cohen, I. G. (2018). Machine learning in medicine: Addressing ethical challenges. *PLoS Medicine*, 15(11), Article e1002689. <https://doi.org/10.1371/journal.pmed.1002689>

Wang, Q., Mao, Z., Wang, B., & Guo, L. (2017). Knowledge graph embedding: A survey of approaches and applications. *IEEE Transactions on Knowledge and Data Engineering*, 29(12), 2724–2743.
<https://doi.org/10.1109/tkde.2017.2754499>

Webber, J. (2012). A programmatic introduction to neo4j. In *Proceedings of the 3rd Annual Conference on Systems, Programming, and Applications: Software for Humanity* (pp. 217–218). Association for Computing Machinery. <https://doi.org/10.1145/2384716.2384777>

Wright, N., Nagle, F., & Greenstein, S. M. (2020). *Open source software and global entrepreneurship*. Harvard Business School Technology & Operations Mgt. Unit Working Paper No. 20-139.
<https://doi.org/10.2139/ssrn.3636502>

©2020 Mayank Kejriwal. This article is licensed under a Creative Commons Attribution (CC BY 4.0) [International license](#), except where otherwise indicated with respect to particular material included in the article.