Google's knowledge vault, sometimes also referred to as Google knowledge graph, is a probabilistic knowledge base that obtains and stores vast amounts of knowledge about the world. This knowledge base can include facts or information about anything including people, places, and other entities. These entities each have unique properties that must be correctly extracted from existing knowledge repositories. However, this usually produces noisy results, so Google's knowledge vault leverages these existing priors with supervised machine learning methods to precisely fuse together relevant information automatically. This allows the average user to quickly obtain relevant, clean information about nearly anything with a simple web search (Dong et al. 1).

The knowledge vault stores information in RDF triples. Each RDF triple has an associated confidence score that represents the probability the knowledge vault believes the triple is correct. The knowledge vault uses a different method than most other knowledge base systems where it avoids storing redundant information about the same entity. As previously mentioned, Google's knowledge vault combines noisy information collected from the web along with prior knowledge obtained from existing knowledge bases, such as Freebase, which is owned by Google, to store the most accurate results. Google has extractors that can automatically crawl the web in search of this noisy information to fill its knowledge base with. Supervised machine learning is used to then verify the information these extractors obtain. They use prior knowledge on an entity based on related facts to make an educated inference. For example, if an extractor returns that Barack Obama was born in Kenya, the knowledge vault can then use its prior models to infer that this information is not true. With this knowledge, it also has the ability to deduce spam-filled websites with misleading information from legitimate sources (Dong et al. 2).

A big difference between Google's knowledge vault and Microsoft's Bing's knowledge graph, or other competitors, is that Google's knowledge vault is holds much more information. As of 2014, according to Google themselves, it held over 1.6 billion triples. Of those 1.6 billion results, 324 million had confidence values of 70% or higher, while 271 million had confidence values of 90% or higher. This makes Google's knowledge vault 38 times larger than the next largest comparable system, which had a mere 7 million confident facts (Dong et al. 2). Those numbers are, of course, much larger today and Google's search engine popularity is reflective of that.

Another difference is that Google's knowledge vault is constantly evolving and being tested for quality. Different extraction methods and prior models are compared to each other and evaluated, and sometimes multiple extraction methods are used together to maximize the best possible information. However, Google's knowledge vault is very secretive, and it's not clear exactly what sources are used when creating these triples. We only know that it comes from a "huge number of Web sources" (Dong et al. 2). The user simply sees all the relevant information for that entity when viewing the knowledge panel. In contrast, Microsoft's Bing's knowledge panel will actually show the web sources of that entity's information directly to the user near the bottom of the panel. It seems like Bing's user interface is more friendly and actually offers even more relevant information at a quick glance than Google's knowledge panel, although this is purely an empirical observation and depends on the data available for the entity a user is searching for.

Google's knowledge vault uses standard methods for relational extraction from text, however it is done at a much larger scale than any other system. The first step in reading, understanding, and obtaining relevant data for the knowledge vault is to use natural language

processing (NLP) tools over each document the extractors obtain. These NLP tools mostly perform entity recognition, which can include many NLP techniques including speech tagging, dependency parsing, co-reference resolution, as well as entity linkage. This helps map out each noun in a document with any other references or corresponding entities. This is done through Google's own in-house built entity linkage system. The knowledge vault has trained for over 4469 predicates, significantly more than any other machine reading system. These can be used to find information of interest to the user. These predicates, or relations connecting entities, can be features or patterns that are stored in these triple (Dong et al. 3).

The knowledge vault can obtain triples from different sources of text data like documents, but it can also get triples from DOM elements on the webpage such as HTML trees, HTML tables and human annotated pages. Google's knowledge vault also uses the path ranking algorithm to perform link prediction over entities connected by some predicate. These links can be analyzed to determine whether two entities have some sort of relation and, if so, through what other entities are they also connected. Different priors can also be fused together, as previously mentioned. These priors can be combined with the feature vector, which contains the vector of confidence values from each prior system and also whether the prior was able to predict or not. Combining the two prior methods can help performance because they each complement each other with their own strengths (Dong et al. 6).

Google's knowledge vault system is much more complex system than can be described in the scope of this review. It can be said that it utilizes many advanced NLP and machine learning techniques to analyze vast amounts of text data into usable information. This information is then indexed similar to any other search engine and is easily read by users. However, it is not a perfect system and there is room for improvement. Google's knowledge vault system currently

looks at each fact as a binary variable, where it's either true or false. But, in reality, triples are typically correlated with each other (Dong et al. 8). Another case is that Google's knowledge vault doesn't always hold all information during all time periods. A fact that is true now may not be true sometime in the future, and the system must be able to understand and store proper information for all time periods and not just the current. Yet another issue is that Google's knowledge vault may have trouble representing all entities. There are always new entities and relations that must be captured, but sometimes this is not possible. The schema for these entities must be expanded to represent these new facts, which is an ongoing challenge (Dong et al. 9).

Conclusively, Google's knowledge vault enables users to quickly obtain lots of relevant information about a particular entity in the knowledge panel. It is more advanced than any previously created system as we discussed in the differences previously. Not only is it highly advanced, but it's also the largest knowledge base currently and it's growing even larger every single day. This is all thanks to advanced NLP and machine learning techniques automating most of the process of sourcing, extracting, organizing, and presenting that data to users in an easy to consume way. The triples within Google's knowledge vault all have associated probabilities which helps it distinguish facts with high confidence from less probable facts. This way the user ultimately gets the best possible information at a quick glance simply by searching for whatever they're interested in right on Google (Dong et al. 9).

Dong, Xin Luna., et al. “Knowledge Vault: A Web-Scale Approach to Probabilistic Knowledge Fusion –.” *Google Research*, 2014, research.google/pubs/pub45634.