Elementary, dear Watson!

Manohar Kaul
IIT Hyderabad
mkaul@iith.ac.in

1. MOTIVATION

In “The Sign of Four", Sherlock Holmes[1] makes some very startling *inferences* about Watson’s whereabouts by informing Watson that he was at the *Wigmore Street Post Office* earlier that morning. On further inquiry by Watson, Holmes explains that he combined (joined) the *observation* that “Watson’s shoes had a *reddish mud* on their instep", with the *fact* that “just opposite the Wigmore Post Office the pavement was freshly dug up to expose a similar reddish mud and it was so positioned that it would be challenging to enter the office without treading into the mud". Contrary to the popular belief that this is *deduction*, this method of reasoning is actually an example of *abductive inference*. Abduction begins from the *facts observed* and then seeks the *simplest hypothesis* (“Occam’s razor"[2]) that *best explains* the facts, while deduction finds data to support a hypothesis.

Given the big data challenge that we presently face, is it then possible to utilize an abductive model (*effect to cause reasoning*) to find the best explanation for the observations, as opposed to the traditional method of forming hypotheses and testing them with observations (*cause to effect reasoning*)? Can our databases *extend* our understanding of this data automatically by inferring explanations from incomplete observations?

2. PROPOSED FRAMEWORK

We will now attempt to mimic Sherlock’s incredible feat using modern day technology.

Observations / Facts: Initially, we know two facts, namely: 1) Watson’s shoes have a red mud on them which is an anomaly and 2) his initial location (home). Using modern systems, Fact 1 could use latest image anomaly detection techniques and Fact 2 could have been Watson’s spatial location recorded and reported from a sensor.

Fact Expander: In his memoirs, Sherlock admits that his inferences are based on collecting as many facts as possible and then *incrementally* expanding them to generate more facts. Can we achieve something similar with modern database systems? A possibility would be to build a *Knowledge Graph (KG)*, as shown in Figure 1, where we start to expand from a central node by querying online repositories like *WordNet* for synonyms and find co-occurring words from *Google n-grams*. For example, “shoe" may co-occur with “walking" or “running". Given such features and their values, how do we pose *meaningful queries* that will lead to more facts? How do we even begin to *intelligently join* data from both online data repositories and also offline databases. A promising approach might be to traverse the *KG* in: 1) *depth-first* and 2) *breadth-first* fashion, while combining the facts intelligently to expand the *KG* with new facts. Could we use a “process of elimination" where we can rule out “sneakers" and “clogs" because the shoes are made of “leather", which most probably implies that Watson might be “walking" (highlighted in blue, Figure 1)? In Sherlock’s own words: *when you have excluded the impossible, whatever remains, however improbable, must be the truth.* Furthermore, can we combine the facts regarding 1) Watson’s initial position, 2) red mud, and 3) walking to pose a spatial query to find locations with red mud (assuming existence of a soil spatial database) that are within a reasonable walking range from Watson’s initial position (highlighted in red, Figure 1)?

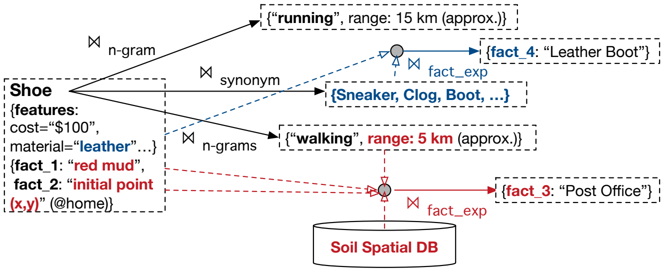

Figure 1: Sherlock Knowledge Graph

Output: A *KG* that shows the facts and all the possible connections that our system could derive with repeated crawls. We can form various hypotheses and test them by following the chain of facts in the resulting *KG*.

3. CONCLUSION

While our example might be a bit contrived, this sort of *exploratory* search (i.e., without issuing the “right" query) might be possible to achieve in certain restricted domains - like the medical domain, where unusual events in sensor data need to be annotated with explanations. The need of the hour is to build data models that are *unbiased* and founded mostly on facts and observations, with little or no priors.

[1] https://en.wikipedia.org/wiki/Sherlock_Holmes

[2] http://goo.gl/8sRktq

CIDR 2017. 8th Biennial Conference on Innovative Data Systems Research (CIDR ’17). January 8-11, 2017, Chaminade, California, USA.