Information Observatories: What Are They Good For?

Grant McKenzie, Krzysztof Janowicz

Department of Geography, The University of California, Santa Barbara, USA

Abstract. Analogies are a fundamental research vehicle that enables humans to explore new domains or scientific problems through the terms and setting of another, more familiar domain. Put simply, scientific analogies serve two purposes. First, they can be used to provide intuitive explanations which are accessible to a broader audience and second, they enable researchers to ask new types of questions that are only possible from within the source domain. Recently, the notion of information observatories have been proposed as an analogy to physical observatories though focused on studying the emerging universe of data rather than the cosmos or a marine ecosystem, for example. Here, we outline why this analogy is useful and present examples of new research questions that arise in this context.

Introduction and Motivation

In 1993, the United Kingdom's science minister, William Waldegrave, offered a prize for the best *layperson's* explanation of the Higgs Boson to better justify the CERN-related spendings to the general public. In response, David Miller introduced his, now famous, analogy of a room full of people chatting among themselves as representation of the Higgs field. In short, Miller states that a famous person crossing the room will attract more guests around her and, due to a crowd forming, will move slowly through the room, while a less prominent person may attract less people and traverse the room faster, an analogy to the mass of particles. This analogy takes a highly abstract and sophisticated target domain and maps it to a source domain that is part of an everyday experience. However, it is important to keep in mind that all analogies are by definition partial mappings, i.e., Higgs analogy neither implies that bosons have genders nor that certain particles are more famous than others. Analogies should not be misunderstood as merely being means of simplification, they also enable us to study a domain in a new light and ask new questions that we may not have envisioned before.

Information Observatories

The notion of a Geographic information observatory (GIO) is such an analogy [1]. They allow citizens and researchers to *observe* the information exposed by myriads of datasets holistically in the same way that a visitor to an astronomical observatory might observe the physical universe. While researchers have successfully explored analogies based on galaxies and nebulas to visualize and navigate large amounts of interconnected data, e.g., Wikipedia articles[1], we are interested here in the *concept* of an information observatory which goes beyond data visualization.

This analogy to physical observatories enables us to ask questions about data that we may have otherwise not considered, questions that could potentially be asked on a much larger (and wider) scale. Trends that may have previously appeared to be domain-specific or limited to a very small sample of a dataset may prove to be not so unique when explored on a larger scale. To give a concrete example, social machines often follow a power law distribution. The vast majority of data is contributed by a few users, a small set of annotations is use for a majority of data, a few topics are highly popular while the majority of contents receives very limited attention, and so on. Such an observation can only be made when comparing different systems, datasets, research publications about these datasets, and so forth.

Even more interesting is the flip side of these scenarios. Researchers typically collect a limited range of data from a certain data source by using a particular procedure, e.g., sampling a certain geographic region with a resolution of 10 miles every 12 hours to access user generated contents. The resulting data is then used to extract features, e.g., for pattern mining. The results are reported based on certain performance measures but typically without looking beyond the scope of the selected sample to study whether the extracted features are merely artifacts of the data source used or whether they persist across data sources. Moreover, are they regionally specific or can they be applied to multiple regions? Are they governed by temporal aspects or do they remain invariant over days, weeks, or seasons? If we do not ask these questions, we may end up learning everything about how humans circumvent the various limitations of *Twitter* [2], e.g., the character limit, but nothing about the underlying principles of communication in social networks, the role of geography in forming topics and spreading them, and so on.

The speed at which new platforms for data collection and creation are developed offers an unique opportunity for GIOs as well. For example, the transportation industry has been revolutionized through the rise of the *gig economy* with applications such as *Uber*, *Car2Go* and others completely changing the way the public commutes. The business model for many of these platforms is such that user-specific data such as pick up and drop off locations, route segments, etc. are recorded and monitored in near real-time. The increased availability and accessibility of these data allows both citizens and researchers the opportunity to monitor transportation patterns from a broad range of sources. Not only can we ask general questions related to transportation patterns, but we are now also able to ask questions such as how the introduction of new platforms alter transportation patterns around the world. The interconnectedness of data creates a

[1] See, for example, Wikigalaxy `http://wiki.polyfra.me/#` .

dense network of user data, car sensor data, governmental traffic data, geosocial check-ins to Points Of Interest, and so forth that enable us to study and realize smart cities beyond what is possible by considering the data sources in isolation.

The physical construction of information observatories as a tangible installation also offers new opportunities to inform the general public. Not only would individuals be exposed to new insights on data presented through such an observatory, but they could also learn about features of the data that they may have not previously considered. Topics related to data privacy, security, provenance and credibility would arise from such a GIO structure. Questions such as 'who is contributing the data?' and 'what are their motivations for contributing such data?' could be asked. Social media content such as tweets are often the source on which trend prediction and even human profiling research is investigated. Is the general public aware that the data they choose to share publicly (though often thought of as only being shared with friends and family) is being used as training data for a classifier being developed to expose potential criminals [3]? Provided this knowledge it is imperative that such an installation be transparent in how data is ingested, analyzed, and exposed by an observatory.

In conclusion, analogies, while only ever partial in their mapping, are important tools in our desire to share and understand new concepts. The idea of a geographic information observatory that observe the data universe can be viewed through an analogy to physical observatories that observe the physical universe. These tools enable both citizens and researchers to approach the data universe from a different and unique angle, asking important and diverse questions that may have otherwise gone unasked.

References

1. Janowicz, K., Adams, B., McKenzie, G., Kauppinen, T.: Towards geographic information observatories. In: Proceedings of the Workshop on Geographic Information Observatories (2014)
2. Krikorian, R.: New tweets per second record, and how! `https://blog.twitter.com/2013/new-tweets-per-second-record-and-how` (2013), [Online; accessed 19-July-2015]
3. Sumner, C., Byers, A., Boochever, R., Park, G.J.: Predicting dark triad personality traits from twitter usage and a linguistic analysis of tweets. In: Machine Learning and Applications (ICMLA), 2012 11th International Conference on. vol. 2, pp. 386–393. IEEE (2012)