Semantic Science: machine understandable scientific theories and data

David Poole
http://www.cs.ubc.ca/spider/poole/

October 13, 2007

Abstract

The aim of semantic science is to have scientific data and scientific theories in machine understandable form. Scientific theories make predictions on data. In the semantic science future, whenever someone does a scientific experiment, they publish the data using a formal ontology so that there is semantic interoperability; it can be compared with other data collected by others, and used to compare theories that make prediction on this data. When someone publishes a new theory, they publish it with respect to an ontology so they can test it on all available data about which it makes predictions. We could all see which theories predict the data better. By the use of formal ontologies, we could determine which are competing theories (when they make different predictions for the same data) and which are complementary. Whenever new data is collected, we can determine which theory better predicts the data. Human-made scientific theories can be compared with machine learned theories (of course, most theories are a mix). Imagine now the best theories applied to new cases: we can use the best medical theory to predict the disease a patient has, the best geological theory to predict where landslides will occur or the best economic theory to predict the effect of a policy change.

This paper is preliminary and always under construction. If you have feedback, more references, please email me at poole@cs.ubc.ca.

1 Introduction

The basic ideas of the semantic science initiative are:

- Theories and data should be machine understandable. This means that they are described in a formal language using ontologies [Smith, 2003] to allow semantic interoperability.

- People publish data using the ontologies. The data will be described using the vocabulary specified by an ontology. Part of this data includes what the data is about and how it was generated.

- Scientists publish theories (hypotheses, models; see Section 2.3) that can make predictions on the data. These predictions can be tested on the published data. As part of each theory is information about what data this theory is prepared to make predictions about.

- New data can be used to evaluate and perhaps update all of the theories that make predictions on this data. This can be used to judge the theories as well as find outliers in the data, which can be statistical anomalies, fraudulent data or some new, little understood phenomenon.

- We can use the formal descriptions of competing theories to devise experiments that will distinguish the theories.

- If someone wants to make a prediction for a new case (e.g., a patient in a diagnostic setting, or predicting a landslide), they can use the best theories to make the prediction. They would either use the best theory or theories, or average over all theories weighted by they ability to predict this phenomenon of interest.

- There is no central authority to vet as to what counts as legitimate scientific theories. Each of us can choose to make decisions based on the what ever theories we want; we will now have more theories available and we will be able to tell which theories are better at predicting unseen data.

There view of semantic science complements other views that build semantic infrastructure for scientists to carry out science [De Roure et al., 2005] that we will not cover in this paper.

2 Science

There are many activities that comprise the enterprise of science. The most important of these are: communication, collecting data (usually by carrying out controlled experiments), and making theories that explain phenomenon.

The activity of science can be seen as the cycle of perception: a scientist observes some phenomenon in the world, formulates hypotheses to explain the phenomenon, and uses these hypotheses to make predictions about future observations. The scientist can then manipulate the world (act) which leads to further

observation. Ultimately, semantic science is about carrying all of these activities in a machine understandable way, so that various aspects can be automated.

Semantic science complements the semantic web [Berners-Lee and Fischetti, 1999; Berners-Lee et al., 2001] which is about publishing data and knowledge in machine understandable ways. There is much work on using the facilities of the semantic web to aid science [Hendler, 2003; De Roure et al., 2005], but they have not emphasized the use of ontologies to enable computers to understand the science. In our proposal, semantic science is also about computers creating new scientific knowledge; making scientific theories brings new knowledge that wasn't implicit in the knowledge given to the system. It is not just a matter of finding information or knowledge integration and interoperability, but is about creating of new knowledge, and applying it to new cases. Of course, semantic science isn't as broad as the semantic web; not all of the knowledge on the web qualifies as scientific data. We hope to thrive alongside the semantic web.

The term "science" is meant to be as broad as possible. We can have scientific theories about any natural or artificial phenomenon. We could have scientific theories about traditional disciplines such as earth sciences, physics, chemistry, biology, medicine and psychology but we would also imagine theories as diverse as predicting which companies will be most profitable, predicting where the best parties are, or predicting who will win football games. The only criteria is that a scientific theory must put itself at risk by making predictions about observable phenomenon.

Semantic science has no prior prejudice about the source or the inspiration of theories; as long as the theories are prepared to make predictions about unseen data, they can be included. We are not, a priori, excluding religion, astrology, or other areas that make claim to the truth; if they are prepared to make refutable predictions, we can test how well their predictions fit the available data, and use their predictions for other data.

Semantic science is trying to be broad and bottom-up. It should serve to democratize science in allowing not just the elite to create data and theories. Like scientists themselves, it should be skeptical of all of the information it is presented with.

2.1 Ontologies

An ontology is a formal specification of the meaning of the vocabulary used in an information system [Smith, 2003]. Ontologies are needed so that different information sources can inter-operate at a semantic level.

As with the semantic web, we want to ensure semantic interoperability, so that a computer can determine the relationship between the vocabulary used in

different data sets and different theories. When people publish data or theories, they can publish them using a published ontology. Data sets or theories that adhere to the same ontology, or to ontologies for which there is a mapping between the common concepts, can then inter-operate.

We expect that scientific communities will try to standardize ontologies, being careful to ensure that communities that have overlapping concerns will try to use common vocabulary. We expect that these ontologies will evolve, taking into account the neighbouring ontologies. This will be a long and slow process.

There is a large body of research on science ontologies, including the Open Biomedical Ontologies (OBO) Foundry (`http://obofoundry.org/`), and the Semantic Web for Earth and Environmental Terminology (SWEET), (`http://sweet.jpl.nasa.gov/`).

2.2 Data

Scientist produce data, and lots of it. Whenever an experiment is carried out or observations are made, data is produced. To be most useful, data should be accompanied with an ontology that specifies the meaning of the data, and with meta-information that specifies information such as how it was collected, what the controls were, what was observed and what the protocol for collecting the data was (from which we may be able to infer expected errors).

Publishing data that adheres to ontologies can be very useful in its own right [Fox et al., 2006]. There has been some work on publishing scientific data such as Community Data Portal (`http://cdp.ucar.edu/`) and the Virtual Solar-Terrestrial Observatory (`http://vsto.hao.ucar.edu/index.php`).

Published data will differ from current data repositories such as the ICI machine learning data repository [Newman et al., 1998] in a number of aspects:

- The people who publish the data are interested in the semantic content of the data: they are interested in what the data says about the world. These are not just abstract data sets to test machine learning algorithms.
- The data sets are not isolated, but overlap. A theory will not be built to cover one data set, but to make predictions the data from many data sets.
- The data will be heterogeneous. Different data sets will describe the same or similar phenomenon at various levels of abstraction and detail. They will have varying control conditions and varying levels of measurement accuracy.
- A data set may contain systematic errors that the author was not aware of

or chose to ignore. It is even possible that some data sets may be fictitious; science is not immune to fraud.

2.3 Theories

Scientific theories and hypotheses make predictions on data. Scientists use the terms "theories", "hypotheses", and "models" to mean similar ideas. Theories, hypotheses and models are usually distinguished by their level of complexity, their generality, and their level of acceptance. We use the term scientific theories in this manifesto as most users will only use the hypothesis or hypotheses that best fits the data for any new prediction; although there may be many narrow or poorly performing hypotheses, they will not get accessed very much. Scientist themselves will create hypotheses that may be untested, and perhaps don't deserve to be called theories, but because we will be able to tell how complex they are, how much data they make predictions about, and how well their predictions fit the data, each person can use their own criteria to decide whether they deserve to be called theories (some won't be very good theories).

In the semantic science future, when scientist publish theories, they publish them in machine readable form and adhere to the ontologies of their disciplines. The theories will made predictions on data that also adheres to the ontology (or for which there is a mapping between the ontologies). We can use these predictions to judge the theories and also to make predictions on new cases.

As part of publishing a theory, a scientist will declare its generality (what it can make predictions about). Scientists will have to be careful about their claims of generality or narrowness. The more narrow a theory, the less it will be used. The more general, the more open to scrutiny it will be. There is a disincentive for a publisher of a theory to make wider claims than they can substantiate; they will be open to more scrutiny as they will be expected to fit more data. There is a disincentive to make theories narrower than necessary: they will be less useful in real applications. Most theories will make no predictions for nearly all of the data sets.

Instead of fitting to a few data sets, theories will be expected to fit to all of the available data about which the theory makes a prediction. Note that the notion of falsifiability of Popper [1959] is too strong for this endeavour. Theories will need caveats, for example probabilities, or allow the hypothesis that something weird is happening for cases where the data is noisy, there are unrecorded exogenous variables, or when the data is mislabeled.

Given multiple theories, we should be able to determine where the theories make different predictions about the same case, as well as when theories only make predictions about different cases. Note that, even in the same domain, and

on the same data, theories can make different predictions but not be competing. For example, consider theories about the retail price of gasoline. One theory may make predictions about how the price changes before weekends, and another may make predictions about how the price changes over the seasons. Determining when theories make incompatible predictions can be the basis of experimental design; an experiment can be designed to create data that will distinguish between the competing theories.

When a scientist has the results of an experiment, they can use it to judge all theories that make predictions about the outcome of the experiment. This means that there can be much more value obtained from each piece of data than when an experiment only compares a couple of theories.

It is important that the science infrastructure treats machine-learned theories on the same footing as human-generated theories. For a long time to come, all theories will end up being some mix of human engineering and learning. For the foreseeable future, humans will design the architecture and provide prior models, and machines will fit parameters to data.

3 Machine Learning

Machine learning will play an important role in this endeavour as computers will do much of the work in generating theories. We want to use the best standards for evaluating theories.

There are many challenges for machine learning, not the least being the ability to make predictions based on heterogeneous data sources. These will be large and diverse data sources at multiple levels of abstraction and detail. The only thing that can be relied on if that the data adheres to an ontology (without this, it is difficult to imagine how multiple data sets could be combined to get the semantic interoperability we need). It will need to combine observational data (which will be vast and varied) with data from controlled experiments (which will be typically smaller, but provide information not available from observation alone).

4 Expert Systems

This should provide the tools for a new wave of expert systems. Practitioners will be able to use multiple theories to make predictions about each new case. They will be able to use the best theories, as judged by all existing data.

They can see is there really is consensus about a case. Note that the practitioners will have to be careful in determining consensus, as they have to make

sure that the theories really are independent evidence. I would expect that many of the theories are variants of each other. How to compensate for this will be a research problem.

Theories can make all sorts of predictions; e.g., probabilistic predictions, that some proportion of the values are in some range, or qualitative predictions. The creators of the theories will specify the meaning of their prediction, but each user of the theories gets to judge the theories based, presumable, on prior plausibility and how well the theory fits the data. Practitioners can ignore good theories at their own peril.

5 Challenges

There are numerous challenges still to be carried out.

The most developed work is on the use of semantics from ontologies to organize and publish scientific data [Fox et al., 2006].

The next major challenge is how to represent and reason with scientific theories that can make predictions on unseen cases. Once this is done, there is challenge of learning these theories: that is, actually doing the science.

Here are some open challenges (in no particular order):

- Making predictions about data sets that are at various levels of abstraction (using more or less abstract terms) and detail (in terms of how they are described in terms of parts and sub-parts).
- Finding suitable representations of theories that can make predictions on varied data sets.
- Determining what predictions a theory can make on a data set that contains extra information or at a different level of abstraction or detail than the scientist anticipated.
- Building an inverse web (like Google) that, given a theory, finds what data this theory makes a prediction about and, given some data, finds the theories that make predictions about this data.
- Use the published theories for experimental design: designing possible experiments that will distinguish the theories.
- A new generation of expert systems that can use the best theories for a new case.
- Building the infrastructure while ontologies are being developed and subject to change.

References

Berners-Lee, T. and Fischetti, M. [1999], *Weaving the Web: The original design and ultimate destiny of the World Wide Web, by its inventor*, Harper Collins, San Francisco, CA.

Berners-Lee, T., Hendler, J. and Lassila, O. [2001], 'The semantic web: A new form of web content that is meaningful to computers will unleash a revolution of new possibilities', *Scientific American* pp. 28–37.

De Roure, D., Jennings, N. R. and Shadbolt, N. R. [2005], 'The semantic grid: Past, present and future', *Proceedings of the IEEE* **93**(3), 669–681. `http://www.semanticgrid.org/documents/semgrid2004/semgrid2004.pdf`

Fox, P., McGuinness, D., Middleton, D., Cinquini, L., Darnell, J., Garcia, J., West, P., Benedict, J. and Solomon, S. [2006], Semantically-enabled large-scale science data repositories, *in* '5th International Semantic Web Conference (ISWC06)', Vol. 4273 of *Lecture Notes in Computer Science*, Springer-Verlag, pp. 792–805.

Hendler, J. [2003], 'Science and the semantic web', *Science* **299**(5606), 520 – 521. `http://www.sciencemag.org/cgi/content/full/299/5606/520?ijkey=1BUgJQXW4nU7Q&keytype=ref&siteid=sci`

Newman, D., Hettich, S., Blake, C. and Merz, C. [1998], 'UCI repository of machine learning databases', University of California, Irvine, Dept. of Information and Computer Sciences. `http://www.ics.uci.edu/~mlearn/MLRepository.html`

Popper, K. [1959], *The Logic of Scientific Discovery*, Basic Books, New York, NY.

Smith, B. [2003], Ontology, *in* L. Floridi, ed., 'Blackwell Guide to the Philosophy of Computing and Information', Oxford: Blackwell, pp. 155—166. `http://ontology.buffalo.edu/smith/articles/ontology_pic.pdf`