UNSW SYDNEY

Sponsored by

capsifi

INTELLIGENT BUSINESS TRANSFORMATION

Creating a Simple Model for Analytics

Introduction

This is a first in a series of exercises whose goal is to create an ontology for helping users manage knowledge about data analytics. We will start with defining two simple concepts: Variables and Models.

Users In the domain of data analytics use the concept "*variable*" to symbolize or describe different attributes. According to Wikipedia:

"In science and research, attribute is a characteristic of an object (person, thing, etc.).Attributes are closely related to variables. A variable is a logical set of attributes. Variables can "vary" - for example, a variable called temperature can be high or low. How high, or how low, is determined by the value of the attribute.

While an attribute is often intuitive, the variable is the operationalized way in which the attribute is represented for further data processing. In data processing data are often represented by a combination of items and multiple variables" ~ Wikipedia

To explain the connections or dependencies between variables analysts use different **models**. These models can be static or dynamic depending on the context of application.

If we take a very simple example, we can have a model that explains the connection between birth year of a person and his age using a simple mathematical model as below.

Age = Current Year – Year of Birth

Here Age, Current Year and Year of Birth, all three can be considered as variables and we have a mathematical model to represent relationship among them. The model used to explain relationships between variables can also be based on statistics, machine learning etc. and not necessarily be represented by an equation. Hence for the simplicity, we treat any model as a black box.

There are two main categories of variables in mathematical and statistical modelling, **dependent** and **independent variables**. The independent variables represent inputs or causes, i.e. potential reasons for variation. Models can be used to test or explain the effects that the independent variables have on the dependent variables. Also, a model can generate the value of a dependent variable, using the independent variables as an input. In our example of

deciding the age of person, Age is the dependent variable, computed using Current Year and Year of Birth as independent variables.

First Semantic Model

Consider a relationship between any two variables X (Independent variable) and Y (Dependent Variable), and we can model that through a model M. M will take X as an input and provide Y as an output.

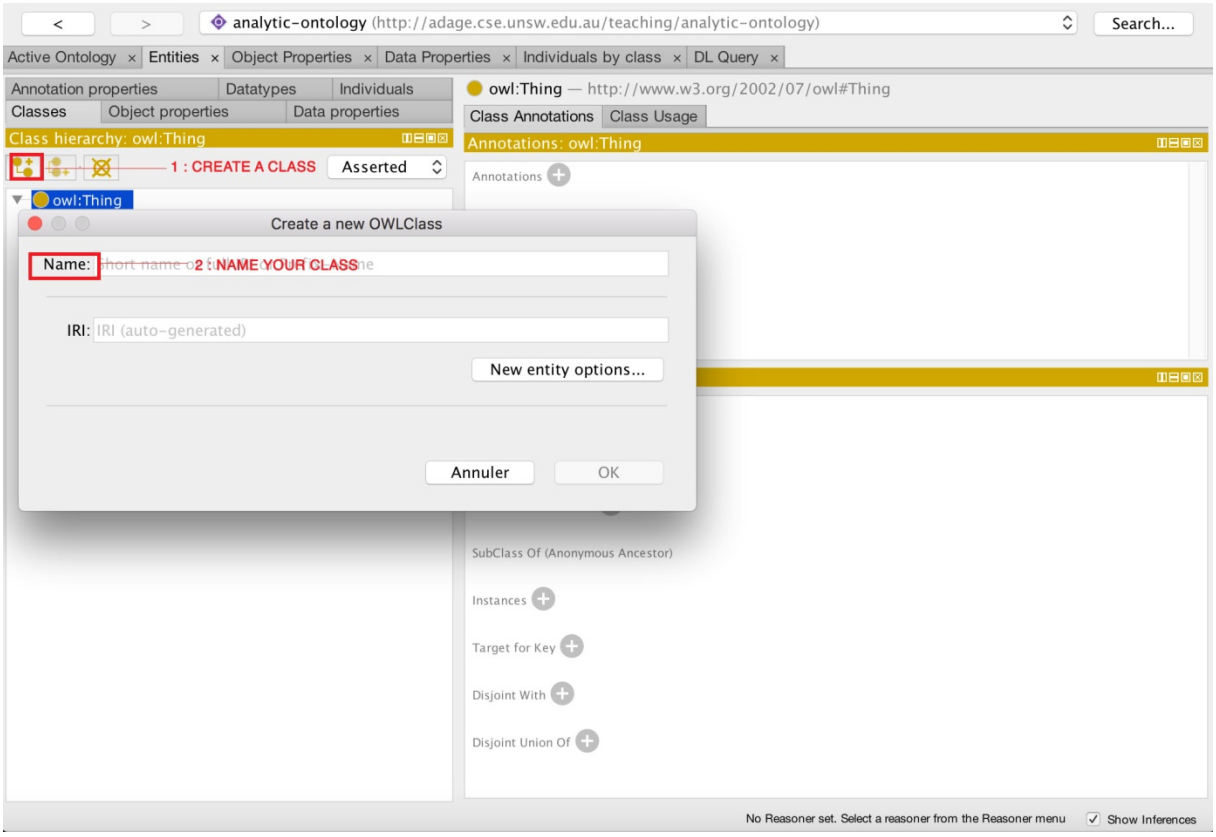
How can we represent this relationship in a notation which is interpretable by human as well as machines? For this we can use a semantic model to represent the abstract concepts of models and variables as classes. Through attributes and relations, we can represent their interactions.

Then to represent specific instances of variables and models, we can create instances of the ontology based classes. This will provide a sound view about any model and its related variables to a user, on which they can query on and develop analytic applications (e.g- to predict one variable given other variables).

Basic Concepts in Creating a Protégé Model

In this step, we will create two classes to represent variable, and model, in order to understand how to create a simple model.

In Protégé, we can create a new ontology, give it a suitable IRI. In the « Entities » tab, we can click on « add class » and name the class as we wish.

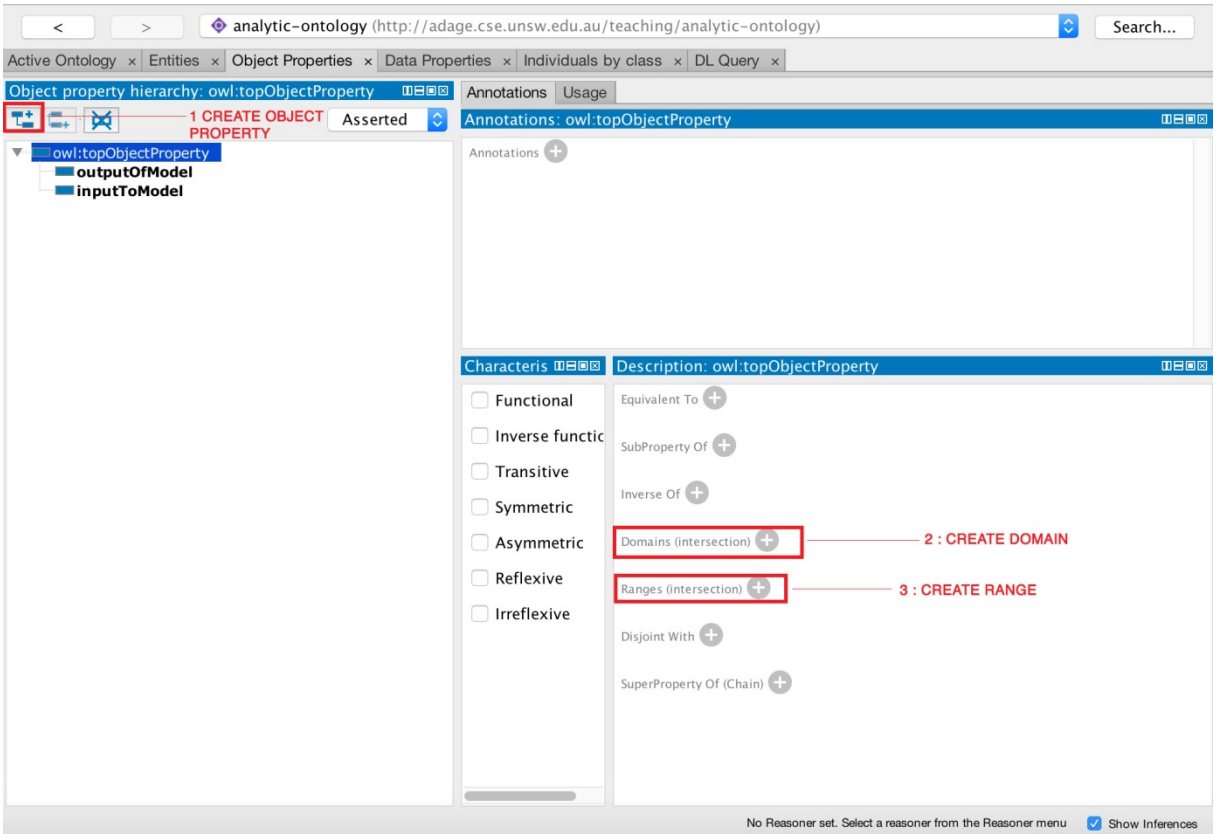


Once we create two classes, we can create relationships between those two classes to represent dependent and independent variables of a model.

We go to the « Object Properties » tab, and click on « add property » and name the property as accordingly.

We will want to define those properties in order to have the right relationship between those two classes. So, we have to define the « Domains » and the « Ranges » of the properties.

We can click on the cross next to Domains, and the select the domain we want, and do the same with the range.



Now you have two classes, linked by a relationship.

Exercise

1) In a new ontology, create two classes: One named “Variable”, the other named “Model”

2) Create two objects property: One named “dependentVariable”, the other “independentVariable”

3) Link the two classes by relationships “dependentVariable” which has “Model” as the domain and “Variable” as the Range. Similarly, “independentVariable” relationship should have “Model” as the domain and “Variable” as the Range.

You can find a sample ontology related to this exercise in the file "Simple Model.owl". Open the file in Protégé and check it.

4) Save the file in Turtle format (call it Simple Model.ttl). Open the file in a text editor (like Notepad) and inspect its content. Try to read the file in text editor and draw the graph that corresponds to this ontology.