Hypothesis generation to explain a dataset.

A form of hypothesis generation that we have implemented takes a dataset with observational data and a semantic description specifying the variable types in the data, and hypothesizes models that may explain the observational data.
The observational data is provided in a CSV file in the "Data" folder in the ASKE_KG_Models project on Eclipse. For illustration, we use a small dataset (file called hypothesis.csv) that includes values for Altitude, AirSpeed, and TotalTemperature. The semantic description of the dataset is:

```
Doc4 is a table
   [double altitude (alias "Altitude") (Altitude {"ft"}),
    double u0 (AirSpeed {"mph"}),
    double tt1 (TotalTemperature {"R"})]
 with data located at "file://Data/hypothesis.csv".
```

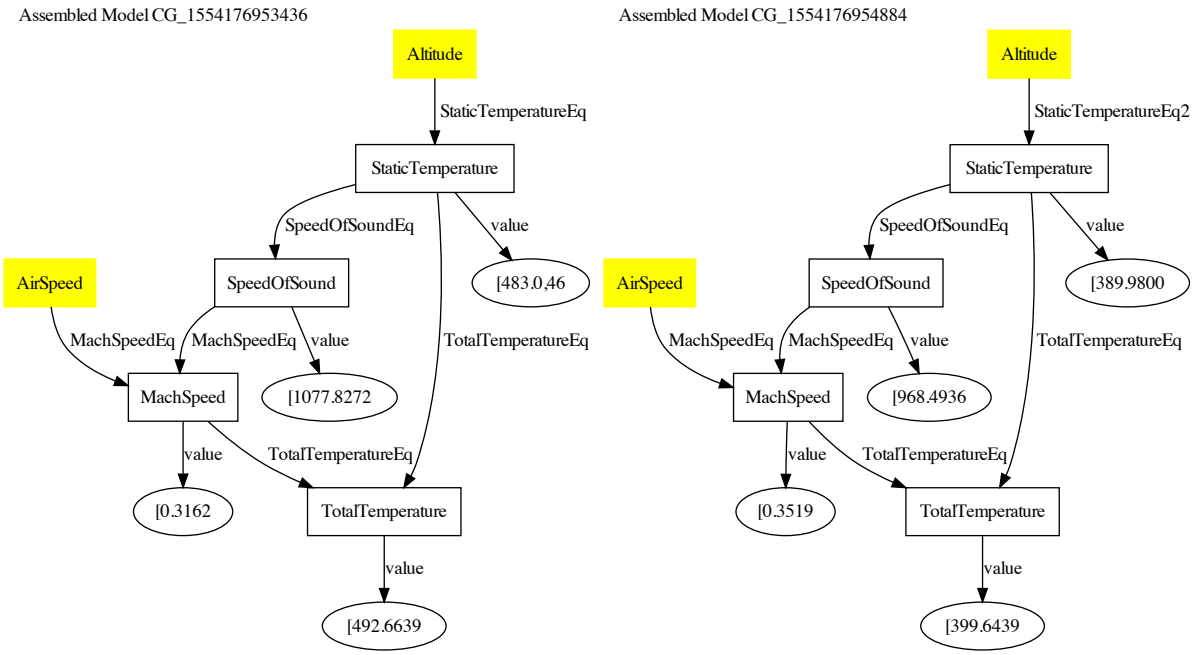
The description specifies for each column in the dataset a semantic variable type (e.g., TotalTemperature). However, it does not specify which variables are inputs and which are outputs. This will be determined by the system once it assembles a model that includes the variable types in the dataset. The description uses "Doc4" as the ID for the document and specifies the location of the CSV file with the data.

The query simply asks for models of the dataset:

```
Find a model for Doc4.
CM: "'Model','ModelAccuracy','Variable','Mean','StdDev','VarError'
'CG_1554176953436','-24.109945943565577', 'StaticTemperature',' [483.0,465.20000000000397]',
'[0.0,3.926123990413682e-12]',null
'CG_1554176953436','-
24.109945943565577','SpeedOfSound','[1077.8272588870623,1057.7802418271908]',
'[6.828041722458576e-13,4.552027814972384e-12]',null
'CG_1554176953436','-24.109945943565577','MachSpeed','[0.31629288283272944,0.3867447064281087]',
'[2.8894707809883298e-15,3.2784380015059896e-15]',null
'CG_1554176953436','-24.109945943565577',
'TotalTemperature','[492.66397873478115,479.1161293780877]','[1.2518076491174055e-
12,4.608928162659539e-12]','[-7.336021265218847,-40.88387062191231]'

'Model','ModelAccuracy','Variable','Mean','StdDev','VarError'
'CG_1554176954884','-108.2299459435655', 'StaticTemperature',
'[389.9800000000026,389.9800000000026]','[2.560515645921966e-12,2.560515645921966e-12]',null
'CG_1554176954884','-108.2299459435655', 'SpeedOfSound', '[968.4936220750214,968.4936220750214]',
'[2.84501738435774e-12,2.84501738435774e-12]',null
'CG_1554176954884','-108.2299459435655', 'MachSpeed','[0.3519993143358906,0.4223991772030636]',
'[8.890679326117937e-16,4.1119391883295465e-15]',null
'CG_1554176954884','-108.2299459435655', 'TotalTemperature', '[399.6439787347827,
403.8961293780863]','[2.84501738435774e-12,3.3002201658549783e-12]','[-100.3560212652173,
-116.1038706219137]'".
```

The model returns two hypotheses (two possible models) because we ran the query after adding a second equation for Static Temperature as in the other demos. The system also generates two visualizations:



For the hypothesis generation task, the DBN execution framework takes the found model and the observational data and computes values for the output variable (in this example, Total Temperature) using the observational data for input variables. Once it has computed those values, it computes the error between the computed value and the observed value of the input variable for each data point in the dataset (last value in the results rows). The DBN framework also computes a form of “model accuracy” using the error in the output variable. In the example, since the values of Altitude in the data set are below 32000, the first model which uses the equation of Static Temperature for lower altitudes has better accuracy (closer to zero).