Inferential Data Modelling in a Query-Answering System

Thomas Fletcher, Alan Bundy, Kwabena Nuamah
T.Fletcher-6@sms.ed.ac.uk, A.Bundy@ed.ac.uk, K.Nuamah@ed.ac.uk

1. Introduction: FRANK

Functional Reasoning for Acquiring Novel Knowledge (FRANK)[1] is a "Third wave of AI" query answering system which performs inferential and statistical reasoning on data from publicly available knowledge bases. It uses a graph-based inference algorithm which decomposes queries into sub-queries until retrievable data is found, and then processes the nodes upwards aggregating children towards the root.
It can perform various aggregation and comparison functions and simple linear regression.

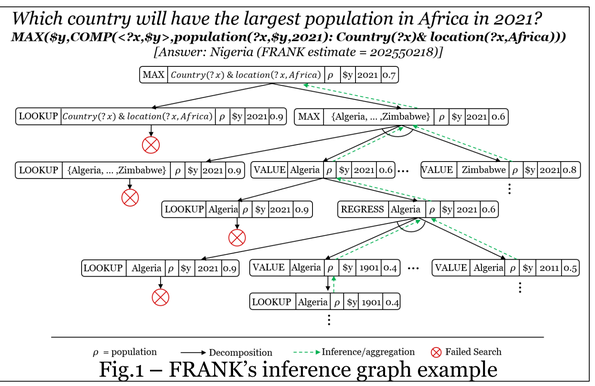

Fig.1 – FRANK's inference graph example

2. Goal System: SMART FRANK

A version of FRANK with extended capabilities, so that it possesses:

- The ability to recognise the various kinds of statistical processes implied by queries along with the statistical nature of data which is retrieved to answer them
- A catalogue of statistical modelling methods ranging from very generic to very specific, and large enough to match (automatically) many types of query and data
- Multiple types of numerical outputs (e.g. specific statistics) and new non-numerical ones, such as specific plots and text descriptions

Some new query type examples

Multivariate Modelling	How does birth rate in European cities vary over population density, country GDP and time?
Analysis of Variance	Taking GDP into account, does life expectancy vary between Italy, Japan and the UK?
Statistical Significance	Is rainfall related to population growth in Asia?
Functional Shape	How does temperature in Swizerland behave over time? | Is ... periodic/linear/exponential/...

Some new output type examples

Text Descriptions	"... rainfall has a linear trend and a periodic component with a period of ..."
Hypotheses P-Values	"... X is different from Y with statistical significance confidence C ..."
Specific Statistics	Correlation, autocorrelation, variance, percentiles, ...
Various Plots	Scatter plots with fits, boxplots, pairplots, high-dimensional visualisations, ...

3. Expert System: SMART

Statistical Methodology Advisor at Reasoning Time (SMART) is an expert system designed to select and perform appropriate statistical methods given a dataset and specific "tags" to begin reasoning from.
It is designed to interface with FRANK by providing it with packages containing appropriate functions to be called at its different processing stages (taggers, an overall decision-making engine and statistical methods). SMART's "expertness" is defined by an ontology graph which contains various input tags, methods, output types and all the entities on which choices have to be made in the process; a simple computational construct (GSM) uses this ontology to make said choices. The core layer of most of SMART's methods is run in R (e.g. Generalised Linear Mixed Models, Analysis of Variance, ...), but an exception is GPy-ABCD[4], a method used primarily for queries requiring functional shape description of the data.

References

[1] Nuamah, Kwabena (2018): Functional inferences over heterogeneous data. PhD. University of Edinburgh.
[2] https://github.com/T-Flet/Graph-State-Machine
[3] Lloyd, James Robert; Duvenaud, David Kristjanson; Grosse, Roger Baker; Tenenbaum, Joshua B.; Ghahramani, Zoubin. "Automatic construction and natural-language description of nonparametric regression models". National Conference on Artificial Intelligence. 2014.
[4] https://github.com/T-Flet/GPy-ABCD

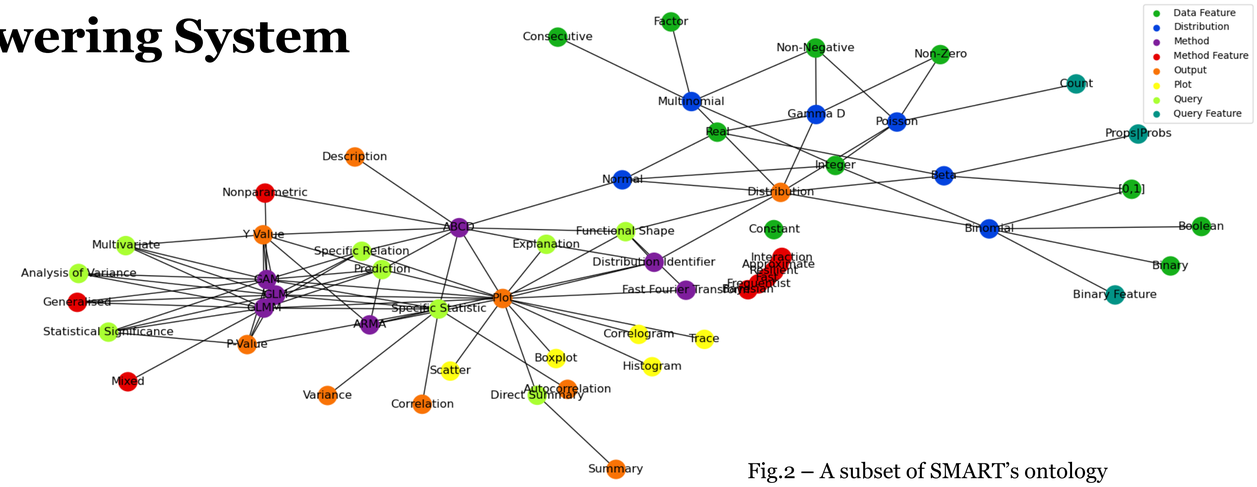

Fig.2 – A subset of SMART's ontology

4. Ontology-Guided Reasoning: GSM

Graph-State-Machine (GSM)[2] is a Python library which implements a computational construct similar to a Turing machine over a graph, where states are node combinations (though more information may be stored) and where the transition function can update both state and graph.
Given a graph with typed nodes (e.g. Fig.2) and a state object from which a list of nodes can be extracted (by an optional `Selector` function), the construct applies two arbitrary functions to perform a step:

- `Scanner` – A generalised neighbourhood function, which scans the graph "around" the state nodes, possibly by type, and returns a scored list of nodes for further processing
- `Updater` - A function to process the scan result and thus update the state and possibly the graph itself

SMART uses a simple GSM to perform its method and output selections and all intermediate choices.

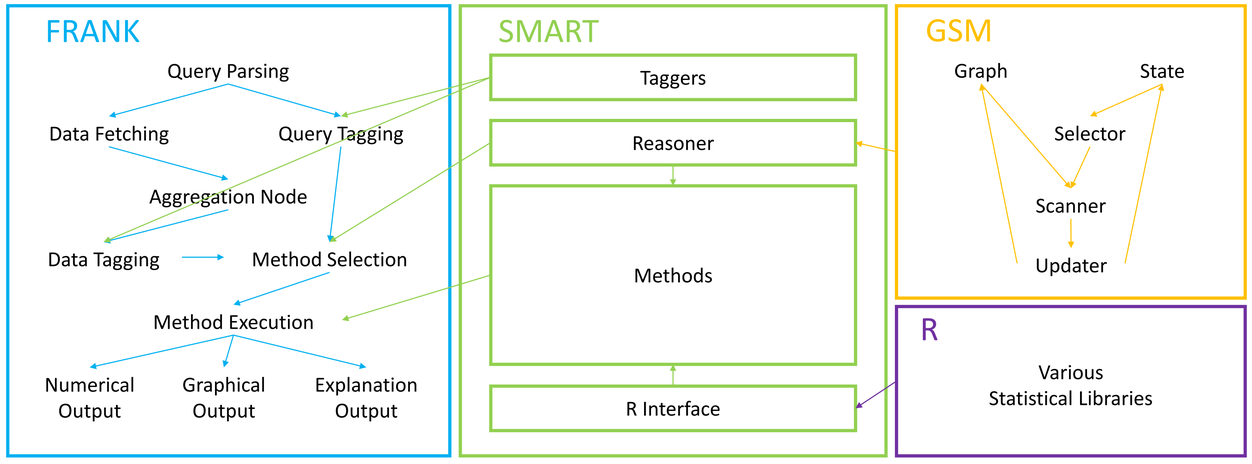

Fig.3 - Component interactions within SMART FRANK

5. A Description Method: GPy-ABCD

Gaussian Processes (GPs) are a very flexible class of nonparametric models which are able to fit data with very few assumptions, namely just the type of correlation (kernel) the data is expected to display. Automatic Bayesian Covariance Discovery (ABCD)[3] is an iterative modular Gaussian Process regression framework aimed at removing the requirement for even this initial correlation form assumption. GPy-ABCD[4] is a new implementation of an ABCD system built for ease of use and configurability, also available as a SMART method.

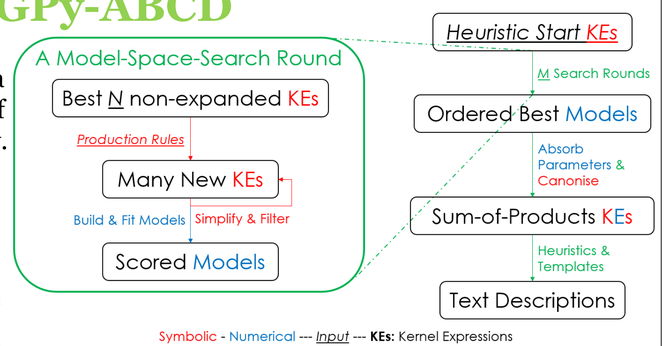

Fig.4 – GPy-ABCD's kernel-space search algorithm