Dimensional Informatics for Scientific Data

Mohamed A. Kamaralzaman, J. Goulding , T. Brailsford

University of Nottingham , School of Computer Science, Horizon Digital Economy Research Institute

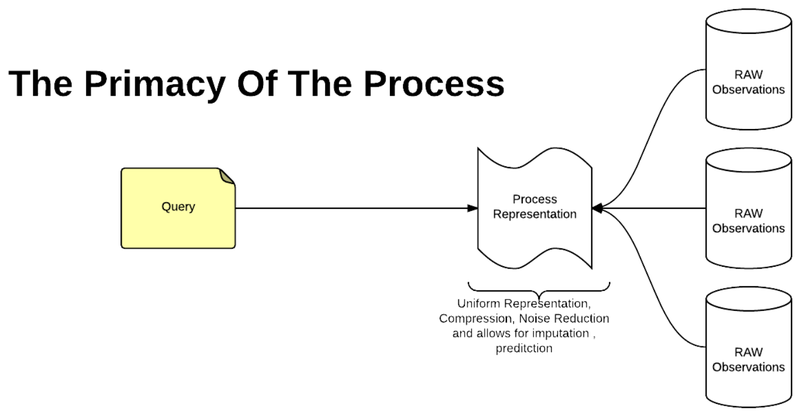
Abstract

As the challenges we face dealing with data increase; manifested in the form of growing volume and expanding variety, a natural evolutionary step has been occurring shifting the focus from managing the `data record' to representing the underlying process that yields it. Modelling processes allows us to overcome the problems of sparsity, irregularity and missing information common when we only record raw observations. This is especially true in the scientific data domain. Abstract representation of processes allow for compression, performing predictions and running what-if scenarios that are not possible when merely storing and retrieving raw observations. In that wake it is clear that improving the techniques with which we represent these processes as well as frameworks allowing for the integration and linking of these process representations into a queryable web of models will allow us to leverage the potential of our data assets beyond its seeming limitations. This research intends to be an effort in that direction with focus on time-series data due to the latter's importance in the scientific domain as well as its relevance to the umbrella project CropBase

The Primacy of the Process

The notion of the primacy of the data record which governed the design, maturation and evolution of today's data management tools is now receding for the notion of the primacy of the underlying process which produces the observed record. While in the former notion the data-point is the object of interest in the latter it is the abstracted and simplified mathematical representation of the underlying process that matters. The benefit of capturing and representing the process underlying a set of observations instead of just the observations themselves is that it allows us to extract information beyond the limitations of which data-points are available and which are not. A representation of the underlying process allows for the simulation of missing data-points , the isolation of signal from noise, assigning probability values for probabilistic data and presenting data values under future scenarios thus overcoming the aforementioned problems of sparsity, addressing the issue of uncertainty and allowing for predictions. Another problem that the notion of the primacy of the process addresses is data compression.

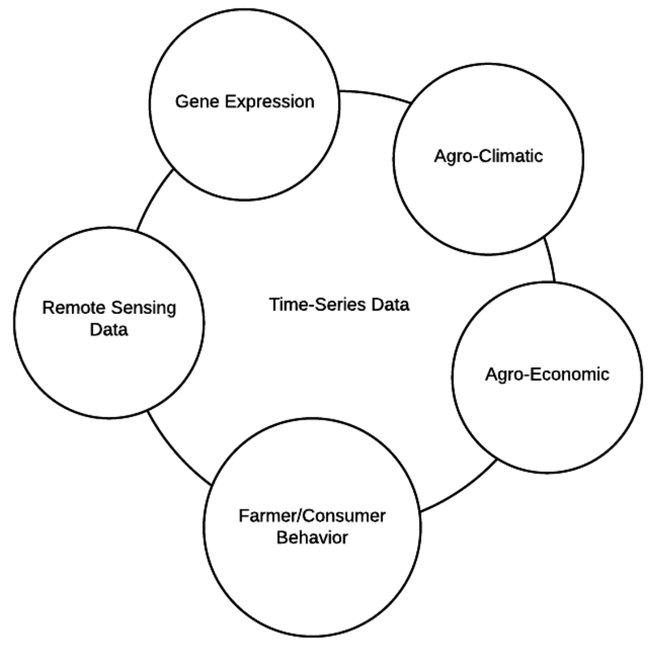
Data Management tools that are in wide use today were designed with the primacy of record in mind not the primacy of the process and therefore they do not address the problem of managing, querying and integrating model representations but rather address a different problem; namely data storage and retrieval.



Time-Series in Model databases

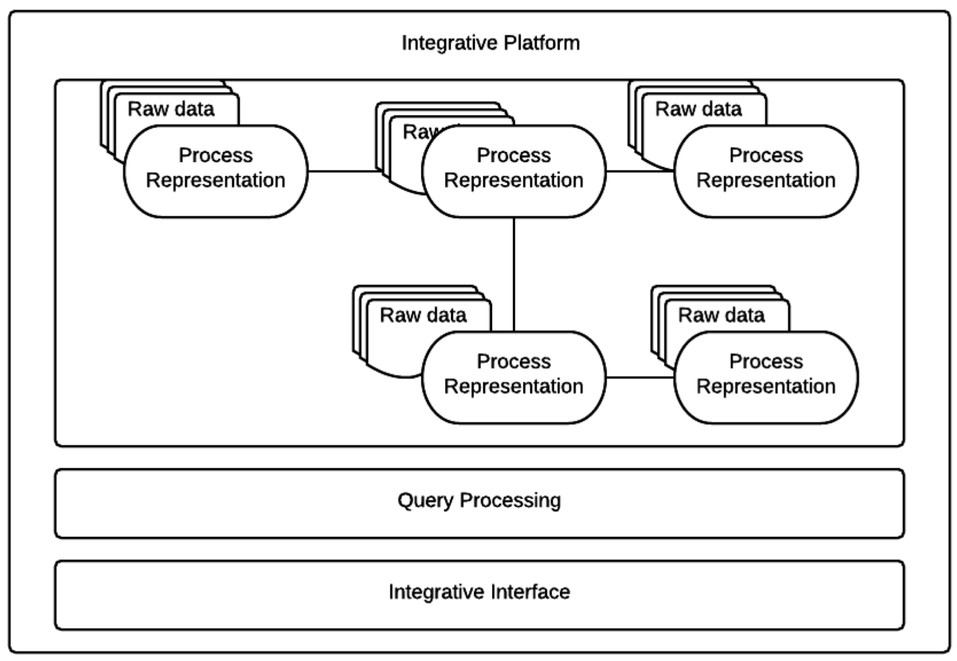
Raw scientific data acquired from lab or field experiments whether manually collected or supplied by wireless sensor networks or recently by drones are never complete nor precise. They may contain outliers and errors, they contain noise and often manifest uncertainty. Typically scientists clean what they acquire and construct simplified representative models for the data describing the underlying processes they are dealing with. This has the advantage of enabling normalized and uniform views of the data as well as enabling predictions and simulations. This is a well known scenario in a wide range of fields an example of which is the scenarios that make use of climatic, agronomic and economic data..

While in general leveraging data observations through model representation is of great and as we have seen established importance; time-series data streams in particular stand out. Research on model databases has generally been more inclined towards static data models - however, much of the experimental research associated with crops arrives in a the form of time series or event processes. Certainly the importance of the temporal dimension for data in real world scenarios is clear, and this is particularly true for crop informatics, climate modelling and economics, all of which are the epicentre of most CropBase relevant scenarios.



A Web of Models

This research approaches integrative Model Based Analytics for Time-Series Data by investigating the best techniques to represent time-series streams as well as event processes. It also extends this objective towards the linking of the resulting models with the aim of constructing a queryable web of process representative models. A consequence is the leveraging the underlying data beyond the limitations of raw observations. Eventually such platform can provide insights that are otherwise more troublesome to achieve such as establishing causation networks or running what-if scenarios.



Acknowledgments

This work is made possible through the support of Crops For the Future