An Ontology for Supporting Knowledge Discovery and Evolution*

Tezira Wanyana[1,2][0000−0002−5139−8421], Deshendran Moodley[1,2][0000−0002−4340−9178], and Thomas Meyer[1,2][0000−0003−2204−6969]

[1] Department of Computer Science, University of Cape Town, South Africa
[2] Center for Artificial Intelligence Research (CAIR), South Africa.
{twanyana, deshen, tmeyer}@cs.uct.ac.za

Abstract. Knowledge Discovery and Evolution (KDE) is of interest to a broad array of researchers from both Philosophy of Science (PoS) and Artificial Intelligence (AI), in particular, Knowledge Representation and Reasoning (KR), Machine Learning and Data Mining (ML-DM) and the Agent Based Systems (ABS) communities. In PoS, Haig recently proposed a so-called broad theory of scientific method that uses abduction for generating theories to explain phenomena. He refers to this method of scientific inquiry as the Abductive Theory of Method (ATOM). In this paper, we analyse ATOM, align it with KR and ML-DM perspectives and propose an algorithm and an ontology for supporting agent based knowledge discovery and evolution based on ATOM. We illustrate the use of the algorithm and the ontology on a use case application for electricity consumption behaviour in residential households.

Keywords: Intelligent agents · ontology · Knowledge discovery · Abductive theory of method.

1 Introduction

In many real world applications, observational data is continuously being captured from complex and erratic physical and social systems that change over time. For example, in earth sciences, natural processes may vary significantly at different locations and typically change over time [22]. In electricity consumption, the household consumption behavior may change depending on the season (summer or winter), day type (week day or weekend) or changes in the demographic characteristics of a given household [27–29]. An integral part of data analysis and scientific inquiry is the detection of phenomena and the development and evolution of theories to analyse and explain these phenomena [11].

Knowledge Discovery and Evolution (KDE) is of interest to a broad array of researchers in Philosophy of Science (PoS) and Artificial Intelligence (AI), in particular, Knowledge Representation and Reasoning (KR), Machine Learning and Data Mining (ML-DM) and the Agent Based Systems communities. While each

[*] Hasso Plattner Institute for Digital Engineering

of these communities have powerful tools and techniques for different aspects of KDE, they each have different perspectives for acquiring information, representing knowledge, revising and updating knowledge, and synthesizing or combining information. AI techniques can be considered as either top down or bottom up. KR is regarded as a top-down AI approach which uses mathematical modelling tools that adopt logic and probability to acquire, represent, and reason about expert knowledge in some domain. ML-DM are regarded as a bottom-up approach and have well established techniques for analysing vast quantities of data and generating complex classification and prediction models. These communities not only have different research cultures and practices which makes collaboration and interaction difficult but also use terminologies in different ways and to mean different things.

In this paper we explore the use of Haig's recently proposed Abductive Theory of Method (ATOM) [11] as a basis to design a unified conceptual model for KDE. We explore ATOM from both KR and ML-DM perspectives and propose an algorithm and an ontology to drive the cognitive loop of a KDE agent. We demonstrate the application and use of the algorithm and the ontology on a use case application for electricity consumption behaviour in residential households.

The rest of the paper is organised as follows. In Section 2, we describe ATOM and how it aligns with aspects of KR and ML-DM. In this section, we also present an algorithm and a unified conceptual model for KDE. In Section 3, we discuss some related ontologies. Section 4 presents a formalization of knowledge discovery and evolution using an ontology. In Section 5, we demonstrate the application of the proposed ontology to the electricity consumption use case and we discuss, conclude and provide some future directions in Section 6.

2 Knowledge Discovery

2.1 Theories of Scientific Method

Scientific inquiry and knowledge discovery are complex processes. Scientists use a plethora of specific research methods and a number of different investigative strategies when studying their domains of interest [11]. Science is a complex human endeavour which articulates aims that it seeks to realize, applies methods in order to facilitate its investigations and produces facts and theories in its quest to obtain an understanding of the world. The scientific method aims to bring some order to these practices.

There are three major types of inference that are applied in scientific inquiry. These are: deduction, induction and abduction. In deduction, the truth of the premises is a guarantee that the conclusion is true. Induction is based on data; for instance the frequency of an occurrence in the given data. It involves generating universal conclusions from specific data or premises. Abduction appeals to explanatory considerations that do not necessarily follow logically from the premises. In an event that there is evidence E and some candidate explanations $H_1 \ldots H_n$ for E, H_i is most likely to be true if it explains E better than any of the other explanations [5].

The inductive and hypothetico-deductive theories are commonly regarded as the two main theories of scientific method. In the inductive theory of scientific method, empirical generalizations are discovered in order to create and justify theories at the same time without having to carry out any empirical testing. On the other hand, the hypothetico-deductive method focuses on the researcher acquiring a hypothesis and testing it by checking its predictive success [11]. Some philosophers of science for example Williamson [30] and Mcmullin [21] argue that abduction is the form of inference that is central to the scientific method. Haig presents a so-called broad theory that incorporates a variety of specific research methods in which the prominent type of inference is abduction [11, 12].

2.2 The Abductive Theory of Method (ATOM)

Haig's Abductive Theory of Method (ATOM) systematically assembles strategies and methods for the detection of empirical phenomena and subsequent construction of explanatory theories. ATOM consists of two overarching methods, i.e. phenomenon detection and theory construction.

ATOM starts with the identification, analysis and extraction of patterns from the data. This would typically comprise of the following steps using statistical and analytical tools: initial data analysis, exploratory data analysis, close replication and constructive replication. This process yields phenomena. These are unexplained "relatively stable, recurrent, general features that researchers aim to explain in the data" [11, 12].

Theory construction is used to provide explanations for the phenomena extracted from the data. ATOM applies abduction in the generation and justification of explanatory theories. Theory construction consists of three sub-methods, i.e. theory generation, theory development and theory appraisal. Plausible theories are generated through abductive or explanatory reasoning using methods like exploratory factor analysis, grounded theory and heuristics. The plausible theories are developed through analogical modeling and are appraised by making judgments on the quality of competing explanations which take into account aspects such as simplicity and consistency with other established theories. The process of theory appraisal applies methods like inference to the best explanation and the theory of explanatory coherence.

In ATOM, phenomena are detected from data and phenomena in turn are used to construct theories. Algorithm 1 shows our interpretation of the basic ATOM process. Note that the steps in lines 12-15 can be repeated several times since theories may emerge from a combination of steps 12 and 13.

We settled on ATOM because, unlike the hypothetico-deductive method, where there is no specific approach to theory formulation, ATOM provides a concrete approach for formulating and generating theories. It also aligns well with both top down and bottom up AI techniques. While ATOM emanated from the behavioural sciences it is applicable to a broad array of complex social, physical and socio-technical systems, such as social networking and health information systems.

Algorithm 1: Basic algorithm for the abductive theory of method (ATOM)

```
input: Data $\mathbf{D}$
output: best explanatory theory t

procedure detectPhenomena(D):
    perform initial data analysis on $\mathbf{D}$ to assess data quality
    repeat until phenomena detected
        suggest pattern using exploratory data analysis
        confirm pattern through close replication e.g. cross validation
        generalize pattern through constructive replication
        if stable pattern found
            $p \leftarrow$ generalised pattern;
    end repeat
    return $p$

procedure constructTheory(p):
    generate plausable theories $T$
    develop theories using analogical modeling
    assess and rank competing theories
    $t \leftarrow$ best theory as explanation for p;
    return $t$

main:
    p=detectPhenomena(D)
    t=constructTheory(p)
return $t$
```

2.3 Machine Learning and Data Mining

Machine learning and data mining are two different areas that have been grouped together in this context. This is because they are both data driven and bottom up and they both offer modern techniques for the detection of phenomena from data which is one of the two main processes in ATOM but note that they are not the same. Knowledge discovery from this perspective involves the discovery of new, previously unknown patterns in the data. Independent examples whose characteristics are different from those defined as normal are first characterised as outliers. Robust techniques have been developed for outlier, anomaly and novelty detection [13, 1, 6], where anomalies are viewed as special kinds of outliers in the data which are of interest to the analyst. Anomaly detection seeks the presence of only one example that cannot be explained by the current model while novelty detection [6] seeks the presence of ”cohesive and representative examples” not explained in the current model. The detected novelty patterns translate to the phenomena for which explanations are sought in ATOM. Where labeled data is not available clustering techniques can be used for detecting and managing patterns.

For applications involving dynamic systems, the machine learning community investigates algorithms to predict the next state of the system. In the simple case of a univariate time series prediction problem, this could involve building a model trained on historical data to predict the next sequence or trend [20] in the data set. Prediction of the next state of the system, sequence or trend in time series or data streams is important for phenomenon detection. It is useful for determining system change or concept drift[7]. For example, when predictions deviate consistently from unexpected observations, then it is possible that the system has changed and that the model needs to be updated.

2.4 Knowledge Representation and Reasoning

One of the benefits of unifying logic and probability which has been a persistent concern in artificial intelligence and philosophy of science is that logic can be used to specify properties that are required to hold in every possible world and probability provides a way to quantify the weight and ratio of the world that is required to satisfy the property in question [3]. KR as a top down approach in this context refers to the tools and techniques that are applicable in the process of theory construction in order to explain the detected phenomena. These are captured formally during the KDE process by the ontology we propose in this paper. Providing explanatory theories, which is one of the overarching steps of ATOM, requires robust techniques for acquisition, maintenance, revision, update of and reasoning about domain knowledge. KR has the ability to provide support for this using tools that apply techniques such as logic and probability. These are applicable in the generation, development and appraisal of theories in order to select the best explanatory theory.

2.5 A Unified Conceptual Model for KDE

An intelligent agent view brings into perspective the aspect of automatic knowledge discovery and evolution. The agent takes in observations in the form of stimuli from its environment. Its role is to deliberate on the observations it has acquired in order to supply appropriate responses based on its beliefs. It also needs a mechanism to represent and communicate the discovered knowledge in a way that is understandable by other software and human agents. This is required in order to attain reproducibility and unambiguous representation of provenance information. Therefore, there is need to settle on a formal ontology that would be required for representing, communicating and reasoning about aspects of observation induced knowledge discovery.

ATOM aligns with the agent's cognitive loop i.e, stimuli/observations, deliberation and response. Kuhn argues that anomalies are a resource that triggers the knowledge discovery process [18]. ATOM can cater for this since it starts with acquiring empirical phenomena from data objects [11]. When the agent acquires observations from its environment, anomalies are detected and temporarily stored. This is done in order to acquire more anomalous observations since a single anomalous example may not be enough to act as evidence for a phenomenon [6]. This is followed by the process of detecting phenomena from the anomalous observations and theory construction to explain the detected phenomena.

In order to perform KDE activities presented in ATOM, an agent based system (ABS) would have to implement tools and techniques from ML-DM and KR to carry out phenomenon detection and theory construction as illustrated in Fig. 1. Unifying the bottom up and top down perspectives of knowledge discovery into an intelligent agent perspective speaks to our desire for interfacing reactive processing with deliberative processing; one of the things we seek to achieve in our unified perspective. Experiential and reactive processing are assumed to be achieved by data driven probabilistic learning methodologies and deliberative processing is assumed to be handled using reasoning methodologies [3].

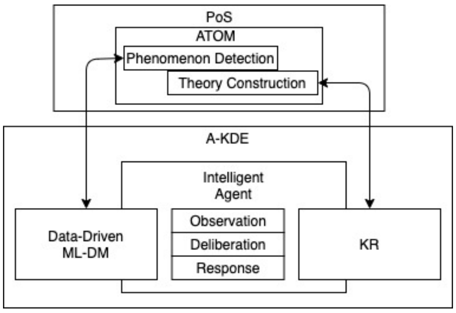


Fig. 1. Knowledge discovery and evolution perspectives

As shown in Fig. 1, ATOM is a unified KDE approach that consists of aspects from both the ML-DM and the KR perspectives. The bottom up (ML-DM) techniques are instrumental in detecting phenomena or novel concepts from the agent's observations and the KR techniques form the foundation of the process of theory construction to explain the detected phenomena in order to fit them into the the body of knowledge that the agent has about the world it inhabits. ATOM forms our stance in this paper in terms of most of the terminology used, evaluation of different ontologies and it also forms the basis of the proposed ontology.

3 Related Ontologies for KDE

In this section, we discuss and evaluate some ontologies for knowledge discovery in terms of the support they offer for observation induced KDE. The evaluation done is in respect to support for capturing aspects of phenomenon detection from data and theory construction to explain the detected phenomena.

Table 1. Comparison of selected ontologies.

ontology has support for:	**LABORS**	**DISK**	**HELO**
Scientific method	Hypothetico - deductive	Hypothetico-deductive	Not explicit
Data	No	Yes, the results produced as a result of executing a workflow from a given line of inquiry	Yes
Phenomena detected from data and phenomenon detection procedure	No	No	Not explicitly stated but records only "Scientific_law" as a statement about phenomena proved by scientific method
Theories that explain phenomena and how they are generated	No	No	Not explicit but records hypothesis and hypotheses set as an explanation and a set of explanations respectively
Models used to elaborate theories	No	No	No
Procedure of theory appraisal	No	No	No
Competing theories	Records research and alternative hypotheses	No	Categorises hypotheses into research, alternatives and negative hypotheses

3.1 LABORS

LABORS (the LABoratory Ontology for Robot Scientists) was designed for representing aspects of scientific experiments for example hypotheses, experimental goals, results, etc. in systems biology and functional genomics. It uses EXPO[3] as an upper level ontology and is used by the robot Scientist [14–17]. The method of discovery that forms the basis of the ontology is hypothetico deductive. The ontology does not include aspects peculiar to knowledge discovery from data or observations and the ontology is very domain specific.

[3] http://expo.sourceforge.net/

3.2 DISK

The DISK (DIscovery of Scientific Knowledge) Ontology [8] for which the requirements are based on the DISK discovery system [9, 10] focuses on representing hypotheses to capture their evolution in automated discovery systems. The hypotheses are supplied by a user or scientist and so there is very little emphasis on the generation of hypotheses or theories. The DISK ontology is constrained to capturing aspects of evolution of user-provided hypotheses.

3.3 HELO

The HELO (HypothEsis and Law Ontology) [26] represents different kinds of scientific statements and links them to their associated probability of being true. It also captures the procedure for obtaining data, research statements and probability statements. The HELO ontology was built from LABORS for the Biomedical domain but it can be used in other domain. Although not explicitly, the HELO ontology provides some support for the KDE ontology requirements as presented in ATOM mainly in the phenomenon detection phase as shown in Table 1. For example, the HELO ontology provides support for *Data.* A concept *scientific_laws* is also used that captures and presents some similarity to *Phenomenon.* The HELO ontology also has a concept similar to *theory* represented as *hypotheses* which captures explanations. However, the nomenclature, taxonomy, interaction and usability of the concepts: *data, phenomenon* and *theory* as represented in the HELO ontology do not capture the aspects and processes in the method that we aim to formalise.

There are other domain specific ontologies designed to support interoperability and reproducibility of scientific investigations and experiments like [24], REPRODUCE-ME [25] for microscopy experiments and OBI (ontology for Biomedical investigations) [2] for biological and medical investigations. Some tools that use ontologies to represent domain knowledge in order to construct theories have also been developed. An example is EIRA (Explaining, Inferring and Reasoning about Anomalies) [23] that was developed for the clinical domain. EIRA [23] does not cater for the representation of hypotheses and their provenance information.

In summary, Table 1 shows a comparison of selected ontologies and the extent to which they offer support for properties of phenomena induced knowledge discovery. The criteria used to evaluate the selected ontologies is extracted from ATOM. The LABORS ontology does not support the criteria used for evaluation in Table 1 because it follows the hypothetico-deductive method specifically in the systems biology and functional genomics domain and focus is on representing aspects of scientific experiments. The same reason applies to the DISK ontology which represents an approach that automates the hypothesise-test-evaluate process that also has characteristics of the hypothetico-deductive method. The HELO ontology, although not explicitly, attempts to capture aspects of the first part of ATOM. However, it does not fully capture the details of the second part. We propose an ontology that caters for the aspects used in Table 1.

4 The KDE Ontology

4.1 Design of the KDE Ontology

The ontology design methodology used was a slight variation of the UPON methodology [4]. UPON is an iterative as well as incremental methodology that is derived from the unified software development process. UPON is use case driven focusing on the development of an ontology that aims to serve either humans or automated systems [4]. The methodology consists of four phases: inception, elaboration, construction and transition phases.

In the *inception phase*, we captured the requirements of the proposed ontology and analysed of existing ontologies. The KDE ontology requirements are mainly based on ATOM [11], a theory of scientific discovery that regards phenomena detected from data as a resource that feeds the knowledge discovery process. The purpose of the ontology is to support an agent based system for knowledge discovery and evolution. The competency questions that the ontology should be able to answer and the household electricity consumption use case were also identified.

A more elaborate analysis was performed in the *elaboration phase* in order to obtain some initial structuring of the main concepts and also to establish the standards to use. We analysed other related ontologies for any reusable concepts and determined how to align the proposed ontology with PROV-O[4], an OWL ontology by W3C provenance working group which provides standards for provenance information.

Design and implementation were the main iterations done in the *construction phase*. Concepts were categorised and the relationships between them established. The ontology was then formalised using OWL (Web ontology language). It incorporates standards provided by PROV [19]. We used Protégé[5], a widely used ontology editing environment to design and implement the ontology.

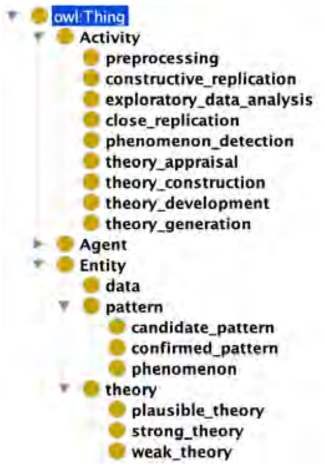
In the *transition phase*, the main activities were around testing the ontology to see if it captures aspects of KDE as presented in ATOM. This involved evaluating the ontology for its support for selected aspects of KDE and checking the ability of the ontology to answer the suggested competency questions.

4.2 The Main KDE Ontology Concepts

In this section , we discuss the main classes represented in the KDE ontology. The ontology is aligned to the W3C PROV standard. The classes, `Entity, Activity` and `Agent` as well as some object properties are drawn from PROV-O. Fig. 2 shows the main classes of the proposed ontology in Protégé. The ontology consists of three main entities and two major activities. The main entities include: `data`, `pattern` and `theory`. The major activities include: `phenomenon_detection` and `theory_construction`. An overview of the main KDE ontology classes and selected object properties is shown in Fig. 3.

[4] https://www.w3.org/TR/prov-o/

[5] https://protege.stanford.edu/



- owl:Thing
  - Activity
    - preprocessing
    - constructive_replication
    - exploratory_data_analysis
    - close_replication
    - phenomenon_detection
    - theory_appraisal
    - theory_construction
    - theory_development
    - theory_generation
  - Agent
  - Entity
    - data
    - pattern
      - candidate_pattern
      - confirmed_pattern
      - phenomenon
    - theory
      - plausible_theory
      - strong_theory
      - weak_theory

Fig. 2. The proposed KDE ontology class hierarchy

Data. Data refers to the instances or information collected for a given purpose. The `data` class captures the required details of the data from which the *pattern* or *phenomenon* is detected. The first activity is to assess the quality of the data using preprocessing (initial data analysis). The activities carried out as part of the preprocessing activity are captured as the class `preprocessing`. The relationship between `preprocessing` and `data` in the ontology is captured using the object property `assesses_quality_of`.

Pattern. The observational evidence required to detect a pattern is provided by data. A pattern is detected from data and this is represented by the object property `was_detected_from` in the KDE ontology. A pattern is an assertion of a recurrent, general feature detected in the data. A phenomenon is a relatively stable pattern, for which an explanation is sought. The phenomenon detection activity captured as `phenomenon_detection` consists of all the tasks undertaken to detect a stable pattern from data. A pattern can be a candidate pattern, a confirmed pattern or a phenomenon captured in the KDE ontology as `candidate_pattern`, `confirmed_pattern` and `phenomenon` respectively. These three types of patterns exhibit a transitive relationship in which a candidate pattern influences a confirmed pattern and in turn, a confirmed pattern influences a stable pattern - the phenomenon. This is captured in our ontology using the `PROV:was_revision_of` object property. A candidate pattern is detected through `exploratory_data_analysis`. This is captured using the object property `was_detected_by`. The candidate pattern is then confirmed through `close_replication`. The object property that captures this relationship is `was_ confirmed_by`. The stability of the confirmed pattern is validated using `constructive_replication`. This is represented using the `was_validated_by`.

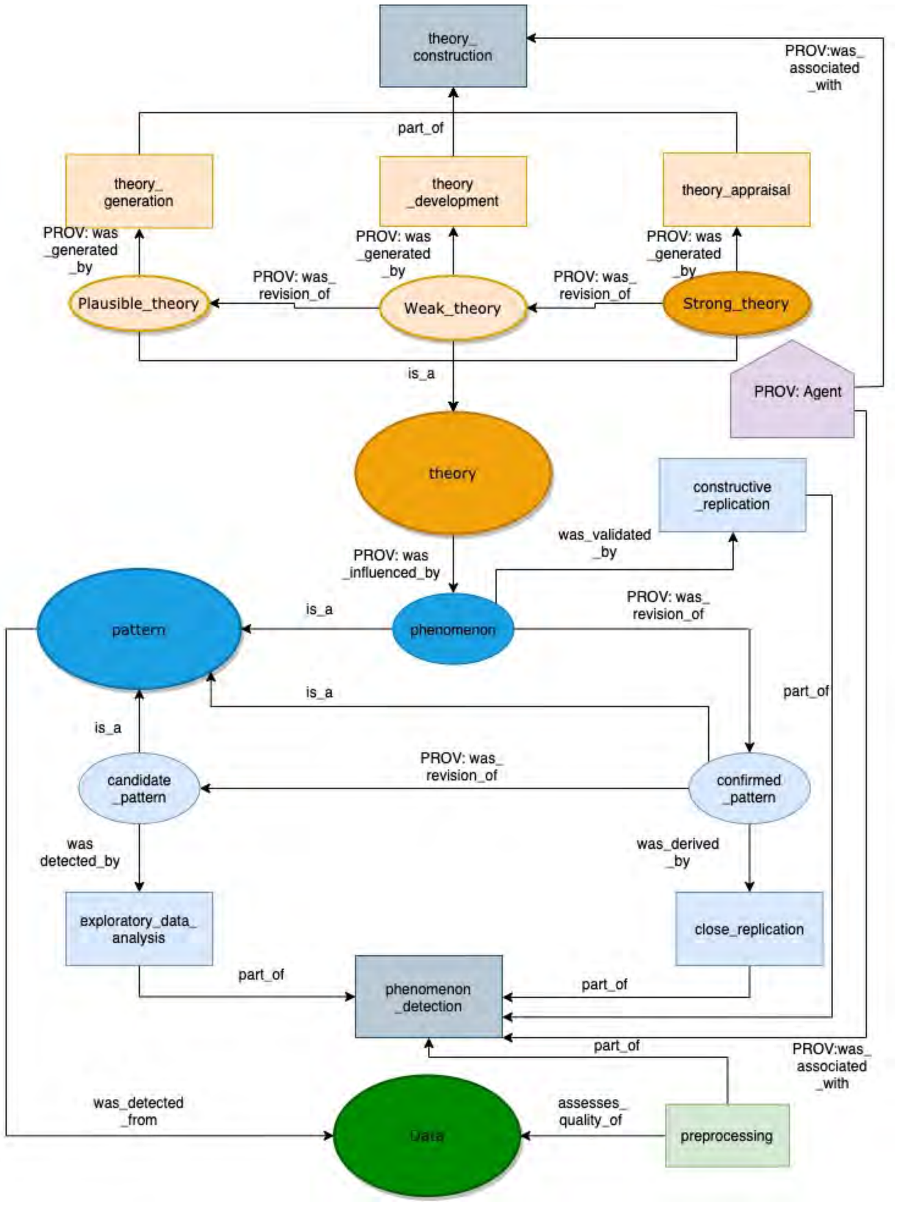


Fig. 3. Overview of the KDE ontology

Theory. This class represents the constructed theories that attempt to explain a given `phenomenon`. The relationship between `theory` and `phenomenon` is represented by the `PROV:was_influenced_by` object property. Theories are only constructed for stable patterns called phenomena. Theories are mainly of three forms; plausible theories, weak theories and strong theories captured as `theory` subclasses: `plausible_theory`, `weak_theory` and `strong_theory`. A weak theory is a revision of a plausible theory and a strong theory is a revision of a weak theory. This relationship is captured in the proposed KDE ontology as `PROV:was_revision_of`. The activity of constructing theories to explain phenomena, captured as `theory_construction` consists of three main subtasks. These include theory generation, theory development and theory appraisal represented as the classes `theory_generation`, `theory_development` and `theory_appraisal`. Theory generation is the process that is used to generate plausible theories. Theory development is used to develop generated theories into weak theories. Theory appraisal represents tasks used to select between competing theories. These aspects as captured using the was the `PROV: was_generated_by`.

4.3 Analysis and Evaluation of the Proposed KDE Ontology

We have presented an ontology that is inspired by the conceptual model for KDE which is based on ATOM. The proposed ontology and the ones briefly discussed in Section 3 formalise the knowledge discovery and evolution process at a meta-level to guide the process of knowledge discovery and evolution. The ontology provides support for the features required for phenomenon detection and theory construction. It also answers the competency questions as required.

The ontology captures features of the data that was used to generate phenomena and the preprocessing that the data was subjected to. Patterns/phenomena detected, at the different stages of stability are recorded along with the techniques used to detect them.

The explanatory theories at each of the levels of theory construction i.e plausible theories, weak theories and strong theories are captured along with the techniques applied during the processes of theory generation, development and appraisal which are all necessary for theory construction.

In conclusion, the proposed KDE ontology captures features, provides support for and distinguishes between data, phenomenon and theory which makes the ontology applicable for agent based KDE from a KR and/ML-DM perspective to record and communicate KDE information. The ontology provides a shared vocabulary and captures the information in a systematic way which aligns with the knowledge discovery process that starts with data or observations, allows for deliberation (phenomenon detection and construction of theories) and return of a response (communicating KDE information). The ontology is accessible online[6].

[6] https://sourceforge.net/projects/akde/

5 Use Case - Household Electricity Consumption

We describe a use case on electricity consumption behaviour of domestic households and use it to illustrate the use of the KDE ontology with two example competency questions answered with SPARQL queries. The use case application is a simplified version of an actual detailed study that used cluster analysis to understand household consumption behaviour in South Africa [27, 28].

Let us assume that the daily electricity consumption of a household h is monitored by collecting hourly data about electricity usage for each day. A 24 element vector is used to represent the consumption daily load profile of the household. Cluster analysis is used to group households with similar daily load profiles for different types of days in the year, depending the season, or whether it is a weekend or weekday. For example there may be different usage on a weekday in summer compared to a week day in winter. The clustering is used to determine the expected consumption behavior of a given household. Each cluster is characterised by the general demographics of the households in it [28, 29]. Demographic data that characterises a cluster could be whether majority of households in the cluster own particular electrical appliances.

Consider two clusters, C_1 and C_2. Let us assume that an agent continuously observes the consumption of household h which it knows to belong to cluster C_1. After some months, through further cluster analysis the agent observes that the daily consumption of household h on summer weekdays increases substantially and now aligns more closely with cluster C_2. This change is identified and captured as a *candidate pattern*. The exploratory method used to generate this pattern and the preprocessing techniques used e.g K-means with unit norm are captured. The agent observes after some time that the increase in consumption for household h persists and it becomes clear that the consumption behaviour now fully aligns with C_2. The *candidate pattern* now becomes and is captured as a *confirmed pattern*. Let us assume that the *confirmed pattern* is then checked for stability by analysing the load consumption of household h again on weekdays during the following year. This *constructive replication* renders the pattern a *phenomenon*. At this point the phenomenon E that the agent seeks to explain is: "h's consumption aligns more closely with cluster C_2 than *C1*"

To provide the best explanatory theory for E, multiple plausible theories $H_1 .... H_n$ may be generated and developed depending on the agent's beliefs. One of the generated explanations H_i could be: "new appliance ownership". In order to develop H_i further, a comparison is made between the demographic characteristics of C_1 and C_2. One of the differences between C_1 and C_2 is that the majority of households in C_2 own appliances, i.e at least a stove, while the households in C1 do not. The agent may find that with the current information, the most probable explanation is new appliance ownership at household h and the appliance is most probably a stove. Given that this is best explanation from all available explanations, this assertion is added to the knowledge base and mark it as a *weak theory*.

Following the electricity consumption behavior use case, two of the answered competency questions are given below.

CQ1(What theories exists for [phenomenon]?) Assuming the phenomenon in question is *'h's consumption behaviour aligns more closely with C2 than C1'* the following SPARQL query in our ontology returns all the theories that explain the phenomenon.

```
SELECT DISTINCT ?theory
     WHERE { ?theory kdeontology:was_Influenced_By
                kdeontology:h’s consumption aligns more closely with
C2 than C1}
```

CQ2 (From what Data was a given pattern detected?)

This competency question would be answered by the following query :

```
SELECT DISTINCT ?Data
                WHERE  kdeontology:h’s consumption aligns more closely
with C2 than C1} kdeontology:was_detected_from
                ?Data
```

6 Discussion and Conclusion

In this paper, we have proposed an ontology that focuses on modelling features of KDE based on an algorithm designed from ATOM. The ontology harmonises vocabulary that would be used by an agent based system that applies ML-DM and KR tools and techniques in tandem to a given use case in order to detect phenomena and construct explanatory theories for the phenomena. This would enable the representation and communication generated knowledge.

Knowledge discovery processes that involve both phenomena detection and theory construction benefit from the full breadth of the ontology. However, in some cases, a few concepts of the ontology may be used. An example is in data streams where the goal is to obtain generalizations from anomalies obtained from continuously acquired data. In such a case a novel pattern is detected as a result of multiple occurrence of unexpected instances. The resulting pattern is a confirmed pattern which may be checked for stability by observing it again in a different setting in order to consider it a phenomenon. In this case the aim is to merely look for stable phenomena and not to explain them.

We demonstrated the application of the algorithms and the KDE ontology on an electricity load consumption use case. We showed how the proposed ontology for KDE represents and captures specific features of theory construction for a phenomenon detected in the consumption behaviour of a given household.

For future work, we intend to extend the proposed ontology to include more features peculiar to automatic KDE and to evaluate the ontology's usability and applicability by a KDE agent.

7 Acknowledgements

This work was financially supported by the Hasso Plattner Institute for digital engineering through the HPI Research school at UCT.

References

1. Aggarwal, C.C.: Outlier analysis. In: Data mining. pp. 237–263. Springer (2015)
2. Bandrowski, A., Brinkman, R., Brochhausen, M., Brush, M.H., Bug, B., Chibucos, M.C., Clancy, K., Courtot, M., Derom, D., Dumontier, M., et al.: The ontology for biomedical investigations. PloS one **11**(4), e0154556 (2016)
3. Belle, V.: Logic, probability and action: A situation calculus perspective. In: International Conference on Scalable Uncertainty Management. pp. 52–67. Springer (2020)
4. De Nicola, A., Missikoff, M., Navigli, R.: A proposal for a unified process for ontology building: Upon. In: Andersen, K.V., Debenham, J., Wagner, R. (eds.) International Conference on Database and Expert Systems Applications. pp. 655–664. Springer Berlin Heidelberg (2005)
5. Douven, I.: Abduction. In: Zalta, E.N. (ed.) The Stanford Encyclopedia of Philosophy. Metaphysics Research Lab, Stanford University, summer 2017 edn. (2017)
6. Faria, E.R., Gonçalves, I.J., de Carvalho, A.C., Gama, J.: Novelty detection in data streams. Artificial Intelligence Review **45**(2), 235–269 (2016)
7. Gama, J.: Knowledge discovery from data streams. CRC Press (2010)
8. Garijo, D., Gil, Y., Ratnakar, V.: The disk hypothesis ontology: Capturing hypothesis evolution for automated discovery. In: K-CAP Workshops. pp. 40–46 (2017)
9. Gil, Y., Garijo, D., Ratnakar, V., Mayani, R., Adusumilli, R., Boyce, H., Mallick, P.: Automated hypothesis testing with large scientific data repositories. In: Proceedings of the Fourth Annual Conference on Advances in Cognitive Systems (ACS). vol. 2, p. 4 (2016)
10. Gil, Y., Garijo, D., Ratnakar, V., Mayani, R., Adusumilli, R., Boyce, H., Srivastava, A., Mallick, P.: Towards continuous scientific data analysis and hypothesis evolution. In: AAAI. pp. 4406–4414 (2017)
11. Haig, B.D.: An abductive theory of scientific method. In: Method Matters in Psychology, pp. 35–64. Springer (2018)
12. Haig, B.D.: The importance of scientific method for psychological science. Psychology, Crime & Law **25**(6), 527–541 (2019)
13. Han, J., Kamber, M., Pei, J.: Data mining: concepts and techniques, waltham, ma. Morgan Kaufman Publishers **10**, 978–1 (2012)
14. King, R.: The adam and eve robot scientists for the automated discovery of scientific knowledge. APS **2017**, X49–001 (2017)
15. King, R.D., Rowland, J., Aubrey, W., Liakata, M., Markham, M., Soldatova, L.N., Whelan, K.E., Clare, A., Young, M., Sparkes, A., et al.: The robot scientist adam. Computer **42**(8), 46–54 (2009)
16. King, R.D., Rowland, J., Oliver, S.G., Young, M., Aubrey, W., Byrne, E., Liakata, M., Markham, M., Pir, P., Soldatova, L.N., et al.: The automation of science. Science **324**(5923), 85–89 (2009)
17. King, R.D., Whelan, K.E., Jones, F.M., Reiser, P.G., Bryant, C.H., Muggleton, S.H., Kell, D.B., Oliver, S.G.: Functional genomic hypothesis generation and experimentation by a robot scientist. Nature **427**(6971), 247–252 (2004)
18. Kuhn, T.S.: The structure of scientific revolutions. University of Chicago press (2012)
19. Lebo, T., Sahoo, S., McGuinness, D., Belhajjame, K., Cheney, J., Corsar, D., Garijo, D., Soiland-Reyes, S., Zednik, S., Zhao, J.: Prov-o: The prov ontology. W3C recommendation **30** (2013)

20. Lin, T., Guo, T., Aberer, K.: Hybrid neural networks for learning the trend in time series. In: Proceedings of the twenty-sixth International Joint Conference on Artificial Intelligence. pp. 2273–2279 (2017)
21. McMullin, E.: The inference that makes science. Zygon® **48**(1), 143–191 (2013)
22. Moodley, D., Simonis, I., Tapamo, J.R.: An architecture for managing knowledge and system dynamism in the worldwide sensor web. International Journal on Semantic Web and Information Systems (IJSWIS) **8**(1), 64–88 (2012)
23. Moss, L., Sleeman, D., Sim, M., Booth, M., Daniel, M., Donaldson, L., Gilhooly, C., Hughes, M., Kinsella, J.: Ontology-driven hypothesis generation to explain anomalous patient responses to treatment. In: Research and Development in Intelligent Systems XXVI, pp. 63–76. Springer (2010)
24. Sahoo, S.S., Valdez, J., Rueschman, M.: Scientific reproducibility in biomedical research: provenance metadata ontology for semantic annotation of study description. In: AMIA Annual Symposium Proceedings. vol. 2016, p. 1070. American Medical Informatics Association (2016)
25. Samuel, S., König-Ries, B.: Reproduce-me: ontology-based data access for reproducibility of microscopy experiments. In: European Semantic Web Conference. pp. 17–20. Springer (2017)
26. Soldatova, L.N., Rzhetsky, A., De Grave, K., King, R.D.: Representation of probabilistic scientific knowledge. In: Journal of Biomedical Semantics. vol. 4, pp. 1–12. BioMed Central (2013)
27. Toussaint, W.: Evaluation of clustering techniques for generating household energy consumption patterns in a developing country. Master's thesis, Faculty of Science, University of Cape Town (2019)
28. Toussaint, W., Moodley, D.: Comparison of clustering techniques for residential load profiles in south africa. In: Davel, M.H., Barnard, E. (eds.) Proceedings of the South African Forum for Artificial Intelligence Research Cape Town, South Africa, 4-6 December, 2019. CEUR Workshop Proceedings, vol. 2540, pp. 117–132. CEUR-WS.org (2019)
29. Toussaint, W., Moodley, D.: Automating cluster analysis to generate customer archetypes for residential energy consumers in south africa. arXiv preprint arXiv:2006.07197 (2020)
30. Williamson, T.: Semantic paradoxes and abductive methodology. In: Armour-Garb, B. (ed.) Reflections on the Liar., pp. 325–346. Oxford University Press Oxford (2017)