Data journeys: knowledge representation and extraction

Enrico Daga a, * and Paul Groth b

a The Open University, United Kingdom
E-mail: enico.daga@open.ac.uk
b University of Amsterdam, Netherlands
E-mail: p.groth@uva.nl

Abstract. Artificial intelligence applications are not built on single simple datasets or trained models. Instead, they are complex data science workflows involving multiple datasets, models, preparation scripts and algorithms. As these workflows increasingly underpin applications, it has become apparent that we need to be able to understand workflows comprehensively and provide explanations at higher levels of abstraction. To tackle this problem, we focus on the extraction and representation of data journeys specifically from data science code. A data journey is a multi-layered semantic representation of data processing activity linked to data science code and assets. We propose an ontology to capture the essential elements of a data journey and an approach to extract such data journeys. Using a corpus of python notebooks from Kaggle, we show that we are able to capture high-level semantic data flow that is more compact than using the code structure itself. Furthermore, we show that introducing an intermediate knowledge graph representation outperforms models that rely only on the code itself. Finally, we reflect on the challenges and opportunities presented by computational data journeys.

Keywords: data science analysis, transparency, data provenance, workflows

1. Introduction

As we increasingly rely on artificial intelligence applications built using complex data science workflows, it has become apparent that we need to understand, integrate, and explain such workflows comprehensively and at a higher level of abstraction. The ACM Principles for Algorithmic Transparency and Accountability [1] emphasise the notions of awareness, audibility, data provenance, and explanation. These principles reflect the idea that stakeholders should have access to and understand what is going on in data science workflows and the AI models that are part of them at the appropriate level of abstraction. Fortunately, the Semantic Web community developed a variety of knowledge representation formalisms to capture fundamental elements of data science workflows, such as provenance, data flows, and high-level activities. Unfortunately, data science workflows are very complex and, although models and techniques for representing code as a data graph exist [2, 3], generating automatically high-level, compact representations is still an open, complex problem.

This article focuses on extracting and representing data journeys, specifically from data science code. Inspired by the social sciences and, in particular, the work of Leonelli [4, 5], we define a data journey as a multi-layered, semantic representation of a data processing activity, linked to the digital assets involved (code, components, data). We propose an ontology for data journeys to capture the essential elements of a data science pipeline and explain it

*Corresponding author. E-mail: enico.daga@open.ac.uk.
in terms of a graph of high-level activities. Similar abstract representations of a data flow have been used to perform
data debugging [6], to extract common motifs in workflows [7], to determine the provenance of data [8], reason on
the relation between policies and process [9], and to support transparency in clinical analysis [10].

Thus, our work aims: (a) to produce a definition of data journey that can be supported computationally, and
a first ontology that reflects such definition; (b) to effectively extract representations of data journeys from data
science pipelines. Our methods do so without the need to execute the pipeline itself, which allows for extensive
scale analysis in, for example, data practices research and empirical software analysis. We show that we can capture
high-level semantic data flow that is more compact than using the code structure itself. Furthermore, we show that
introducing an intermediate graph representation of the data flow outperforms models that rely only on the code
itself.

Specifically, the contributions of the article are as follows:
– a definition of data journey suitable for computational approaches;
– a rich ontology for representing data journeys, mapped to existing, state of the art, relevant ontologies;
– an approach for knowledge extraction from Python-based notebooks to generate a compact, high-level explana-
tion of the data journey;
– experiments demonstrating how a graph-based representation of the data science code outperforms one based
on the code syntax in a machine learning pipeline supporting the extraction process.

The rest of the article is structured as follows. We begin with a discussion of related work (Section 2). Next, we
look at our definition of a data journey and the associated ontology (Section 4). In Section 5, we detail our approach
for knowledge extraction. We report on the application of the method and its evaluation in Section 6. Finally, we
reflect on paths forward for further research around computational data journeys.

2. Related Work

We introduce the background around data journeys and then discuss related work in provenance, workflows, and
methods for capturing the data flow from programs. The larger field of explainable AI has been covered comprehen-
sively by multiple recent surveys [11, 12]. Here, we focus on how AI/data science algorithms and systems can
be understand through the flow of data. As others have argued [13], this is a critical and under-studied area.

Data journeys The broad notion of data journeys have been discussed in the data studies literature. In particular,
with the recent edited volume by Leonelli and Tempini, which brings together different case studies from plant phe-
nomics to climate data processing and studies them through the lens of data journeys [4]. Fundamentally, they argue
that the journey a dataset goes through, its lineage or provenance, is a powerful unit of analysis for understanding it.

Provenance and provenance representations The need to understand the provenance of data has been well docu-
mented in the data management [14] and web [15] literature, which has investigated approaches for representing,
extracting, querying, and analysing provenance information. Indeed, the importance of understanding provenance
for web information led to the W3C Prov standard for provenance interchange [16] as well as the recent Coalition
for Content Provenance and Authenticity1. We refer to the two surveys cited above for more information about
provenance systems. Our work, in particular, builds upon these existing representations in order to provide a multi-
layered view of a data journey allowing different levels of abstraction to sit alongside one another. Specifically, we
build on the notion of datanode as specified in the Datanode Ontology [17], developed to express complex data
pipelines to reason upon the propagation of licences and terms and conditions in distributed applications [18].

Workflow abstractions The need to tie data to the workflow that generated it has been recognised in the scientific
workflow community [19]. An essential contribution of this work is that, for workflow analytics and reusability, dif-
ferent granularity levels of workflow representations and associated provenance (e.g. high-level tasks in the domain
vs command-line tool parameters) should be captured [7, 20]. These representations can then be bundled together

1https://c2pa.org
with the corresponding data assets and other documentation, creating a research object [21] that can be published using web standards [22]. However, most data science programs are not expressed with such workflow formalisms.

Capturing dataflow from programs While these approaches work for workflow systems where tasks and their dependencies are systematically defined, and most applications use flexible programming languages. It is the case for data science methods [23]. To perform analytic and assistance, a variety of work has looked at extracting high-level workflow-like representations from code or logging information. Tessera, for example, has looked at extracting high-level tasks from logs of exploratory data analysis [24].

Furthermore, in work most closely related to ours, researchers have investigated extracting provenance representations from programs’ abstract syntax trees (AST). CodeBreaker constructs a machine-interpretable knowledge graph from program code to support end-user tasks such as code search and recommendation [3]. noWorkflow [8] uses the AST of a program to automatically instrument code during execution to capture provenance. Similarly, mlinspect [6] extracts a provenance representation during the execution of python code. It, however, uses a higher-level representation designed for data science code, which allows for richer data dataset debugging capabilities. ProvenanceCurious [25], and Vamsa [26] also use the python AST, but instead of instrumenting at runtime, they use static analysis to infer a provenance graph. We adopt a similar approach of using static analysis. However, unlike these approaches, we use a hybrid method based on knowledge engineering and machine learning to infer high-level semantic types over a resulting graph. Additionally, our approach tackles the need to maintain multiple levels of abstraction.

3. Definition

In this section, we provide a definition of Data Journey. The notion of data journey has been discussed in the data studies literature. Specifically, [4] defined it as the “movement of data from their production site to many other sites in which they are processed, mobilised and re-purposed.”. The work in data studies emphasises the difficulty of understanding data journeys empirically because of a multitude of perspectives. Our definition has the objective of being consistent with the one of [4] but also to relate with the literature from Web semantics, specifically, data provenance [27]. Hence, we introduce a layered semantics perspective to the definition of data journeys:

A Data Journey is a multi-layered, semantic representation of a data processing activity, linked to the digital assets involved (code, components, data).

Thus, a journey is multi-layered, as to allow a multiplicity of perspectives that can be overlaid to describe the process. This multiplicity can help to capture (parts of) the context around a data journey while still allowing for computational analysis to be performed. Hierarchical, because any useful representation needs to be linked to the concrete assets involved, either directly or via intermediate abstractions.

Although our definition is open-ended and allows for multiple (even alternative) perspectives to co-exist, in this work, we conceptualise data journeys in the following layered structure:

- **Resources**: resources used in the data journey such as source code files, software libraries, services, or data sources.
- **Source Code**: human readable and machine executable instructions, for human authoring, such as a Python script.
- **Machine Representation**: any machine interpreted representation of the instructions, such as an Abstract Syntax Tree (AST) or a query execution plan.
- **Datatnode Graph**: as defined in [28], a graph of data-to-data relationships, such as variables, imported libraries, and input and output resources. Such abstraction provides a structure of the data flows, abstracting from issues such as control flow, and focusing on data-to-data dependencies.
- **Activity Graph**: a graph of high-level activities, inspired by the notion of Workflow Motifs [7].

While the first three components pre-exist the data journey, i.e. they do not pertain to the knowledge level [29], the remaining represent two distinct, although interconnected, representation layers. Here, we aim at automatically constructing such a layered representation, to demonstrate that data journeys can be automatically identified from the source code of a data science pipeline.
4. Ontology

We build on previous work and design an ontology as the reference knowledge model of a data journey, satisfying our definition. Our methodology is based on reusing successful, relevant models, completing them with concepts derived from the data used in this work. Specifically, we reuse concepts from three approaches: the W3C Provenance Ontology PROV-O [30], the Workflow Motifs Ontology [7] and the Datanode ontology [17]. Compared to the pre-existing models, DJO provides a unified view of the data journey, linking the two layers (the data flow layer and the activity layer). In addition, it extends the Datanode Ontology by specifying node types.

The Data Journey Ontology (DJO) reuses fundamental concepts from those three ontologies into a new conceptual model with layered semantics. The namespace of the DJO is http://purl.org/datajourneys/, the preferred prefix is djo. The design rationale of DJO is one of layered semantics. The ontology should be able to capture fine-grained data flows but also high-level activities, as super-imposed abstractions. The core components are three classes: Datanode, Activity, and Journey.

Datanode The Datanode class represents a data object. We performed a thematic analysis of 20 randomly chosen notebooks from the dataset collected for the experiments (more about it in Section 6) in order to identify possible data node types, depending on the role they have in the process. The result is as follows:

- **Constant**: a value hard-coded in the source code, for example, the value of an argument of a machine learning instruction.
- **Input**: a pre-existing data object served to the program for consumption and manipulation.
- **Output**: a data node produced by the program.
- **Parameter**: any data node which is not supposed to be modified by the program but is needed to tune the behaviour of the process. For example, the process splits the data source into two parts, 20% for the test set and 80% for the training set. 2, 20%, and 80% are all parameters.
- **Support**: any data node pre-existing the program, which is used without manipulation. Includes:
  - **Reference**: any datanode used as background knowledge by the program, for example, a lookup service or a knowledge graph. Such datanode pre-exists the program and is external to the program.
  - **Capability**: any datanode which provides capabilities to the program, including pre-existing modules, functions, and imported libraries.
  - **Documentation**: any datanode which does not affect the operation of the program, such as source code comments and documentation.
- **Temporary**: any datanode produced and then reused by the program that is not intended to be the final output.

The following groups of classes are mutually disjoint:

- Input, Output, Support, and Temporary
- Capability, Documentation, and Reference

Datanodes are connected to each other in a data flow akin to a provenance graph. Possible relationships subsume the provenance notion of derivation. Table 1 shows the hierarchy of object properties, with references to the notions reused from existing sources, and documented in the ontology with a rdfs:seeAlso relation.

Activity In a Data Journey, sibling datanodes are supposed to be grouped together in activities, in order to provide a more abstract representation of the data journey. The activities included in DJO are mutually disjoint sub-classes of Activity: Analysis, Cleaning, Movement, Preparation, Retrieval, Reuse, and Visualisation. Except for Reuse, the other classes are derived from concepts defined by the Workflow Motifs ontology. Activities instances are connected to each other in a sequence, using the object property previous (and its inverse next). We note that the semantics of these two properties does not imply derivation but merely states the succession of activities in a data journey. Each activity instance is then connected to the involved datanodes through the property includesDatanode (and its inverse inActivity).

Journey Finally, both activities and Datanodes are meant to be connected to one data journey via the functional property inJourney.
Table 1

List of object properties connecting instances of class Datanode in the Data Journey Ontology. Sources: PROV-O (PO), Datanode Ontology (DN), and Workflow Motifs (WM).

<table>
<thead>
<tr>
<th>Property</th>
<th>PO</th>
<th>DN</th>
<th>WM</th>
</tr>
</thead>
<tbody>
<tr>
<td>derivedFrom</td>
<td>✓</td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td>analysedFrom</td>
<td>✓</td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td>cleanedFrom</td>
<td>✓</td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td>computedFrom</td>
<td>✓</td>
<td></td>
<td></td>
</tr>
<tr>
<td>copiedFrom</td>
<td>✓</td>
<td></td>
<td></td>
</tr>
<tr>
<td>movedFrom</td>
<td>✓</td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td>optimizedFrom</td>
<td>✓</td>
<td></td>
<td></td>
</tr>
<tr>
<td>preparedFrom</td>
<td>✓</td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td>augmentedFrom</td>
<td>✓</td>
<td></td>
<td>✓</td>
</tr>
<tr>
<td>combinedFrom</td>
<td>✓</td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td>extractedFrom</td>
<td>✓</td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td>filteredFrom</td>
<td>✓</td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td>formatTransformedFrom</td>
<td>✓</td>
<td></td>
<td></td>
</tr>
<tr>
<td>groupedFrom</td>
<td>✓</td>
<td></td>
<td></td>
</tr>
<tr>
<td>sortedFrom</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>splitFrom</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>refactoredFrom</td>
<td>✓</td>
<td></td>
<td></td>
</tr>
<tr>
<td>remodelledFrom</td>
<td>✓</td>
<td></td>
<td></td>
</tr>
<tr>
<td>retrievedFrom</td>
<td>✓</td>
<td></td>
<td></td>
</tr>
<tr>
<td>visualisedFrom</td>
<td></td>
<td>✓</td>
<td></td>
</tr>
</tbody>
</table>

5. Extracting data journeys

In this section, we describe our approach to extracting data journeys. Specifically, we examine the problem of automatically deriving data journeys from source code, focusing on identifying activities and their composition. We ground our method on the following hypothesis:

– Given a data science program, it is possible to automatically extract a data journey from the code (a) by generating a datanode graph from the code; (b) making use of machine learning techniques to classify high-level activities in such graphs automatically; and (c) collapsing adjacent activity nodes involved in the same activity.

The method assumes source code as input (e.g. a python notebook), then generates a data node graph reusing symbols from the code (variable names, functions, operators, etc...). It then uses this information to train a classifier for identifying activity types. We note that we only focus on the generation of the data node graph structure and on the activity types, leaving the automatic support for the remaining features of the ontology (e.g. data node arc labelling and node types) to future work.

We divide the approach into four steps: (i) data node graph extraction, (ii) knowledge expansion, and (iii) knowledge compression.

In the first step - data node graph extraction - the method traverses the code Abstract Syntax Tree (AST) and transforms the structure into a Datanode graph. In such a graph, nodes represent data elements (such as files, variables, constants, and libraries), while arcs represent relationships, labelled with operations derived from the code – e.g. importing a library or assigning a variable – or with names of functions, method calls, or operators.

In the second step - knowledge expansion - we aggregate the data node graphs into a knowledge graph and derive frequent relationships (arcs in the graph) in a Frequent Activity Table (FAT). The most frequent activities are then annotated with the Data Journey Ontology (limited to Activity types in our experiments). The FAT allows us: (a) to generate automatically a set of rules encoded in SPARQL CONSTRUCT queries, to materialise the annotations,
and (b) to produce a dataset of annotated data node graphs for training a classifier able to assign an Activity to each node in the data node graph. Furthermore, the classifier is applied to derive DJO activities automatically.

In the last step – data journey generation, adjacent nodes with the same activity types are collapsed, producing a summarised semantic representation of the data science pipeline – completing the Data Journey.

In the following sections, we describe each step in the abstract, leaving the details of the execution to the evaluation section, where we apply it to Python notebooks [31]. While we use Python here, the approach applies to any programming language that can be expressed as an AST.

5.1. Datanode graph extraction

We describe deriving a data node graph from a Python notebook. The algorithm implements heuristics, transforming the code structures into a graph. We organise the process in three steps: (a) first, we extract the source code from the notebook into a single python script; (b) second, an algorithm traverses the code and generates a directed, labelled graph from the source, serialised in DOT format; then, (c) the directed graph is re-engineered into RDF, applying namespaces and generating entity names. Listing 1 (in the appendix) illustrates the algorithm which generates the directed graph in pseudo-code. The process starts with traversing the Abstract Syntax Tree (AST) to build a directed graph where the nodes are data objects, such as libraries, files, constant values, and occurrences of variables, and arcs are dependency relationships, in the spirit of provenance models. The algorithm assigns unique identifiers (and human-readable labels) to graph nodes and assigns relationships applying an initial semantic layer derived from the code syntax (e.g. considering elements such as import, assignment operators, or function calls). The output of this step is a directed graph where nodes are data objects and arcs are relationship types derived from the code syntax.

The resulting directed graph is re-engineered into RDF. The direction of the arcs is reversed in the spirit of provenance graphs. In addition, we add namespaces and use an entity function, which returns an RDF resource from the node label, applying the appropriate conversions to valid URI strings. The result is a data node graph, the first layer of our Data Journey.

5.2. Knowledge expansion

The previous section described obtaining a data node representation from the source code. This section will use the data node graph to derive background knowledge for training a Machine Learning algorithm able to annotate nodes with data journeys activities. Background knowledge includes the following components:

– The Data Journeys Ontology (DJO), introduced in Section 4, which specifies the following activity types:
  - Reuse, Movement, Preparation, Analysis, and Visualisation
– A dataset of (frequent) data node arcs, manually annotated with DJO Activities – Frequent Activity Table (FAT)
– A set of rules (encoded as SPARQL construct queries) mapping frequent arcs in the data node graphs to DJ Activities, using the FAT – Frequent Activity Rules (FAR)
– Training dataset for the machine learning model, to predict Activities of unknown data node nodes – Machine Learning Training Dataset (MLTD)

**Frequent Activity Table (FAT)** From a statistical analysis of the data node graphs produced, one can determine the more frequent arcs types than others. Such arcs are manually annotated with activity types. For example, nodes receiving an arc named "dj:importedBy" can be associated to a Reuse activity, or an arc derived from a popular function such as read_csv to a Movement activity. Similarly, a print arc demonstrates a Visualisation activity, and so forth. The objective is to have sufficient coverage of activity types in the table to use the annotation for training a machine learning model able to automate such annotations on all nodes in the data node graphs.

**Frequent Activity Rules (FAR)** From the FAT, we build rules to materialise the annotations. These are encoded as SPARQL CONSTRUCT queries to apply to the dataset of data node graphs.
Machine Learning Training Dataset (MLTD)  We derive a training dataset from the table of annotated data node arcs. The input data includes a set of data nodes annotated with activity types (derived from the FAT). We can then enhance the activity types with background knowledge from existing knowledge models such as CodeBERTa or produce RDF2Vec embeddings directly from the data node graphs.

Machine Learning Application We use the dataset to evaluate the performance of potential ML approaches. Specifically, we evaluate a set of machine learning methods to automatically annotate the missing nodes with activity types in this phase. The most promising method is selected and applied to predict the activities of remaining and less frequent nodes.

5.3. Knowledge compression

We annotated the data node graph with activity types in the knowledge expansion phase. In this phase, we aim at obtaining a summarised view of the data journey. The knowledge compression task has the objective of generating activity instances by analysing sibling data nodes annotated with the same activity type. Activities may span multiple adjacent data nodes. For example, a Reuse activity applies to all imported libraries. Therefore, a better representation would generate a single instance of Reuse’s activity type, grouping all the import data nodes under a shared activity instance. The output is an Activity graph, whose nodes are instances of activities and arcs the previous relation. Crucially, activities may occur multiple times in different parts of the data node graph.

The process applies the components described so far through the algorithm illustrated in Listing 2. The process implements the following recursive pipeline, starting from the root node (the program):

1. Focus on a node and collect the set of previous nodes. Then, group those nodes by activity type, relying on the annotation property hasActivity, produced in the previous phase.
2. If the focus node was already linked to an activity instance in a previous iteration (property inActivity), link previous nodes having the same activity type to that activity instance with an inActivity property.
3. Generate a new activity instance for each remaining group, linking the related nodes with an inActivity relation.
4. Link the new activity instances to the focus node activity instance with a previous relation.
5. Repeat for each previous node.

The resulting graph is, therefore, a compressed view of the data science pipeline in the form of an activity graph. Such graph data is overlaid on the data node graph, completing the task of building a data journey, a layered, semantic representation of the data science program.

6. Evaluation

This section evaluates our hypothesis that it is possible to automatically derive a data journey by applying a combination of methods that abstract the source code into a layered, semantic representation. Specifically, we perform experiments to evaluate the effectiveness of classifying nodes within the program with higher-level semantics as expressed in the Data Journeys Ontology. Crucially, we aim to verify whether the graph structure is helpful in this classification process over just the use of the code itself. We evaluate:

1. the feasibility of automatically generating a graph representation, anchored to the source code,
2. the ability of data node graphs to support the automatic classification of activity types, and
3. the ability of activity graphs to offer a representation that is substantially more compact than the data node graph.

In the remainder of the section, we follow the approach introduced in Section 5, and we apply it to Python notebooks. Our dataset consists of the 1000 most popular python notebooks from Kaggle as of April 2020. Kaggle is one of the main hosts for data science competitions. Data node graphs are constructed for each notebook.

---

3 As measured by the ‘hotness’ parameter of the API.
4 kaggle.com
algorithm extract-directed-graph is
  input: notebook
  output: graph
  graph ← []
  ast ← AST(notebook)
  visit ast components:
  when is import:
    ?var ← the library handler in the code
    ?library ← the python imported module
    graph U { ?var assignedFrom ?library }
  when is Module:
    visit ast components
  when is Class:
    visit ast components
  when is Function:
    for each argument:
      ?arg ← reference to the function argument
      ?var ← handle to the local variable
      graph U { ?var _argToVar ?arg }
  when is Expression:
    if object method call:
      ?source ← object variable before method call
      ?target ← object variable after method call
      ?method ← method name as arc label
      graph U { ?target ?method ?source }
      for each argument:
        ?source ← reference to argument variable
        digraph U { ?target ?method ?source }
    if function call:
      for each argument:
        ?target ← the variable passed as argument
        ?source ← reference to the function argument
        ?method ← function name as arc label
        graph U { ?target ?function ?source }
  when is Assignment:
    for each left-hand variable:
      ?source ← reference to the function argument
      for each right-hand variable:
        if augmented assignment:
          ?operation ← the operation of the augmented assignment
          graph U { ?target ?operation ?source }
          else:
            graph U { ?target assignedFrom ?source }
    for any node without an incoming arc:
      ?notebook ← the passed notebook
      graph U { ?notebook appearsIn ?node }
  return graph

Listing 1 Pseudocode of the algorithm to extract a directed graph from the Python code.
The implementation of the approach and all data assets used in the evaluation are provided as supplementary material to this submission for review is available for auditing and reproducibility [32].

6.1. Datanode graph extraction

In this step, we apply the algorithm described by Listing 1, which generates the directed graph. For example, given the following python code snippet (Kaggle: https://www.kaggle.com/dansbecker/random-forests):

```python
import pandas as pd
# Load data
melbourne_file_path = '../input/melbourne-housing-snapshot/melb_data.csv'
melbourne_data = pd.read_csv(melbourne_file_path)
# Filter rows with missing values
melbourne_data = melbourne_data.dropna(axis=0)
```

the algorithm generates the following directed graph, represented here in DOT format:

```dot
strict digraph "" {
  pandas -> "random-forest.ipynb" [label=importedBy];
  "pd(0)" -> pandas [label=assignedFrom];
  "pd(0)" -> "random-forest.ipynb" [label=appearsIn];
  "../input/melbourne-housing-snapshot/melb_data.csv(0)" -> "random-forest.ipynb" [label=appearsIn];
}```
The arcs of the graph incorporate semantics derived from the code structure, partly assigned from heuristics in the algorithm, and partly derivable from the names of operators used in the code. For example, the relation importedBy links a library to the python script; the relation add links a variable (or constant) to its target variable; the relation print links a processing node to an output node; the relation assignedFrom is applied to local handlers of imported libraries as well as regular variable assignment operations. In addition, the example shows how the implementation takes care of linking multiple occurrences of the same variable together, generating a new node every time the variable value changes (for example, when re-assigned or when a method is called on a variable object, assuming the internal state of the data object is affected by the method call). Finally, the implementation ensures that variables are interpreted in the context of their scope. The number in parenthesis represents the outer scope, while the dollar sign distinguishes different instances of the same variable name. For example, it is distinguishing a variable name used within the context of a function from the same variable name used in the outer scope (e.g. melbourne_data, lines 7-10).

The resulting directed graph is re-engineered into RDF. In particular, the direction of the arcs is reversed. Then, we add links from the notebook entity to each one of the nodes. In addition, we add namespaces and use an entity function which returns an RDF resource from the node label, applying the appropriate conversions to valid URI strings. In the implementation, we use the following namespaces:

```PREFIX dj: <http://purl.org/dj/> .
PREFIX k: <http://purl.org/dj/kaggle/> .
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#> .
PREFIX nb: <http://purl.org/dj/kaggle/random-forest#> .
```

The k: namespace prefix is used for naming notebooks, the lib: namespace prefix is applied to python libraries, while the default namespace is used for any other entity generated. The above code generates the following data node graph in RDF:
The execution of the process produces 804 data node graphs. The remaining notebooks either included syntax errors, they did not include any source code, or the process took more than 10 minutes to complete. We leave the study of possible optimisations to future work.

### 6.2. Knowledge expansion

The output of the previous phase are data node graphs expressed in RDF. Such representation has the benefit to be directly linked to the source code, to enable us to overlay a graph representation on the program code. In this phase, we aim at identifying activities for each one of the node. Data node arcs are labelled according to clues derivable from the code (e.g. import statements and assignments, see Listing 1) or by reusing code elements (e.g. object methods, function names, ...).

**Frequent Activity Table (FAT)** Following our method, we compute statistics of data node arcs and focus on the most frequent relationships in a Frequent Activity Table (FAT). The dataset includes 3384 distinct relations, from the more frequent ones (the most frequent being `assignedFrom`, 90370 occurrences) to relations occurring less frequently. We manually annotated the arcs occurring at least 2000 times in the dataset with data journey activities. The intended meaning is that when a given relation occurs, the target node can be qualified as being the result of the specified activity. Table 2 shows the relations occurring at least 1000 times in the dataset and related annotations. However, we observed that the `Movement` activity was underrepresented in the table; therefore, we selected two more arcs mapped to that activity.

**Frequent Activity Rules (FAR)** From the FAT, rules are derived and encoded as SPARQL CONSTRUCT queries. Listings 3 and 4 show examples of such rules. The rules materialise a new statement using the DJO annotation property `hasActivity`, connecting data node entities to subclasses of `Activity`. We apply the rule sets to the data node graphs and materialise the new triples. An excerpt related to the guide example can be seen in Listing 5.

```
1 PREFIX dj: <http://purl.org/dj/>
2 PREFIX : <http://purl.org/datajourneys/>
3 CONSTRUCT { ?s :hasActivity :Reuse }
4 WHERE { ?s dj:importedBy ?o . }

Listing 3: We can infer from the datanode arc `importedBy` that the target node relates to a Reuse activity

```
```
1 PREFIX dj: <http://purl.org/dj/>
2 PREFIX : <http://purl.org/datajourneys/>
3 CONSTRUCT { ?s :hasActivity :Visualisation }
4 WHERE { ?s dj:print ?o . }

Listing 4: We can infer from the datanode arc `print` that the target node relates to a Visualisation activity

```
Frequent Activity Table (FAT). Relations occurring at least 1000 times in the datanode dataset and related annotations. The table also distinguishes relations explicitly asserted by the AST traversal algorithm and the ones that are implicitly derived from the code.

<table>
<thead>
<tr>
<th>Datanode arc</th>
<th>Occurrences</th>
<th>Asserted</th>
<th>Annotation</th>
</tr>
</thead>
<tbody>
<tr>
<td>print</td>
<td>16888</td>
<td></td>
<td>Visualisation</td>
</tr>
<tr>
<td>iteratorOf</td>
<td>11825</td>
<td>Y</td>
<td>Preparation</td>
</tr>
<tr>
<td>importedBy</td>
<td>10855</td>
<td>Y</td>
<td>Reuse</td>
</tr>
<tr>
<td>_argToVar</td>
<td>9713</td>
<td>Y</td>
<td>Reuse</td>
</tr>
<tr>
<td>Add</td>
<td>7334</td>
<td>+</td>
<td>Preparation</td>
</tr>
<tr>
<td>append</td>
<td>6840</td>
<td></td>
<td>Preparation</td>
</tr>
<tr>
<td>subplots</td>
<td>6128</td>
<td></td>
<td>Visualisation</td>
</tr>
<tr>
<td>apply</td>
<td>5471</td>
<td></td>
<td>Preparation</td>
</tr>
<tr>
<td>Div</td>
<td>5115</td>
<td></td>
<td>Preparation</td>
</tr>
<tr>
<td>fit</td>
<td>4415</td>
<td></td>
<td>Analysis</td>
</tr>
<tr>
<td>read_csv</td>
<td>4412</td>
<td></td>
<td>Movement</td>
</tr>
<tr>
<td>astype</td>
<td>4191</td>
<td></td>
<td>Preparation</td>
</tr>
<tr>
<td>plot</td>
<td>4036</td>
<td></td>
<td>Visualisation</td>
</tr>
<tr>
<td>train_test_split</td>
<td>3667</td>
<td></td>
<td>Preparation</td>
</tr>
<tr>
<td>tanh</td>
<td>3546</td>
<td></td>
<td>Analysis</td>
</tr>
<tr>
<td>DataFrame</td>
<td>3494</td>
<td></td>
<td>Preparation</td>
</tr>
<tr>
<td>title</td>
<td>3243</td>
<td></td>
<td>Preparation</td>
</tr>
<tr>
<td>mean</td>
<td>3200</td>
<td></td>
<td>Preparation</td>
</tr>
<tr>
<td>add</td>
<td>2843</td>
<td></td>
<td>Preparation</td>
</tr>
<tr>
<td>subplot</td>
<td>2545</td>
<td></td>
<td>Visualisation</td>
</tr>
<tr>
<td>drop</td>
<td>2527</td>
<td></td>
<td>Preparation</td>
</tr>
<tr>
<td>predict</td>
<td>2483</td>
<td></td>
<td>Analysis</td>
</tr>
<tr>
<td>merge</td>
<td>2404</td>
<td></td>
<td>Preparation</td>
</tr>
<tr>
<td>map</td>
<td>2345</td>
<td></td>
<td>Preparation</td>
</tr>
<tr>
<td>head</td>
<td>2314</td>
<td></td>
<td>Preparation</td>
</tr>
<tr>
<td>[ ]</td>
<td>1033</td>
<td></td>
<td>Movement</td>
</tr>
<tr>
<td>[ ]</td>
<td>678</td>
<td></td>
<td>Movement</td>
</tr>
</tbody>
</table>

Listing 5: Example of data nodes annotated with activity types.

**Machine Learning Training Dataset (MLTD)** We create a training dataset from the table of annotated data node arcs. The input data includes a set of data nodes annotated with activity types (derived from the FAT). These can be enhanced with background knowledge, such as from existing knowledge models such as BERTcode or producing RDF2VEC embeddings from the data node graphs. Activities in the FAT are not equally represented. For example, Reuse, Preparation, and Visualisation have many more nodes than Analysis and Movement. Therefore, we select four arcs for each activity to give more balanced coverage to the training data. These are highlighted in Table 2.
The resulting dataset includes the following information: the notebook, the graph node, the arc used to derive the annotation, and the annotated activity. The data includes 802 notebooks and 29282 nodes. Statistics of arcs, activities, and number of nodes in the training dataset are reported in Table 3.

Table 3
ML Training Dataset, statistic of activity types and nodes.

<table>
<thead>
<tr>
<th>Arc</th>
<th>Activity</th>
<th>Nodes</th>
</tr>
</thead>
<tbody>
<tr>
<td>fit</td>
<td>:Analysis</td>
<td>1316</td>
</tr>
<tr>
<td>tanh</td>
<td>:Analysis</td>
<td>299</td>
</tr>
<tr>
<td>predict</td>
<td>:Analysis</td>
<td>959</td>
</tr>
<tr>
<td>read_csv</td>
<td>:Movement</td>
<td>1727</td>
</tr>
<tr>
<td>copy</td>
<td>:Movement</td>
<td>431</td>
</tr>
<tr>
<td>to_csv</td>
<td>:Movement</td>
<td>490</td>
</tr>
<tr>
<td>Add</td>
<td>:Preparation</td>
<td>2078</td>
</tr>
<tr>
<td>append</td>
<td>:Preparation</td>
<td>2111</td>
</tr>
<tr>
<td>iteratorOf</td>
<td>:Preparation</td>
<td>4499</td>
</tr>
<tr>
<td>importedBy</td>
<td>:Reuse</td>
<td>1622</td>
</tr>
<tr>
<td>_argToVar</td>
<td>:Reuse</td>
<td>9709</td>
</tr>
<tr>
<td>plot</td>
<td>:Visualisation</td>
<td>1408</td>
</tr>
<tr>
<td>subplots</td>
<td>:Visualisation</td>
<td>1705</td>
</tr>
<tr>
<td>print</td>
<td>:Visualisation</td>
<td>6405</td>
</tr>
</tbody>
</table>

Machine Learning Application
In this phase, we train an ML model to annotate the missing nodes with activity types automatically. In the next section, we will report on experiments with multiple machine learning algorithms. We select the most promising method and use it to predict the activities of remaining non-frequent nodes. The output of the learned classifier is a data node graph whose nodes are all annotated with activity types, using the DJO OWL annotation property :hasActivity.

6.3. ML classification experiments

We train classifiers that, given a node representing an Activity from a notebook, predict its high-level semantic type - one of Analysis, Movement, Preparation, Reuse, Visualisation. We tested standard classifiers available in scikit-learn, specifically Logistic Regression, Decision Trees, Gaussian Naive Bayes, Linear Support Vector Machine and a Multi-layer Perceptron.

We develop tests using two different methods for representing nodes as input to the classifiers:

1. CODEBERTa - In this setting, we embed the code string associated with a node using CodeBERTa a pre-trained transformed based language model trained on a large corpus of the programs from the CodeSearchNet [34]. The corpus used contains roughly 6 million functions spanning six programming languages.

2. RDF2Vec - In this setting, we use RDF2Vec [35] to embed each node based on its structural position in a knowledge graph constructed from the RDF representation of the notebooks used during the experiment. RDF2Vec is set to use the following parameters: the node is represented based on its context, i.e. a continuous bag of words (CBOW). The context is determined by extracting paths in a random walk around the node. A maximum of 100 paths are extracted with a maximum depth of 10.

For each setting, we employ two testing regimes:

R1 Nodes are randomly split between training and testing sets.

---

5 The data in the table was produced by querying the training dataset (CSV) using SPARQL Anything [33]
6 https://huggingface.co/huggingface/CodeBERTa-small-v1
R2 Nodes are randomly split between training and testing sets but the train and test sets do not share notebooks.

A 70/30 train test split is used. Hyperparameters are as detailed above. R1 simulates the behaviour where one wants to classify other (e.g. long-tail) nodes from within a codebase, which contains some already classified nodes. Whereas R2 aims to simulate an environment in which one is trying to classify nodes from a notebook that has not been seen before. For each classifier, setting and regime combination, we use 10, 100 and 200 notebooks randomly sampled from the corpus. Experiments are repeated 10 times for each combination. Table 4 shows the average knowledge graph size across experimental runs.

The experiments showed an increasing accuracy when providing a larger set of notebooks. Table 5 presents the best results achieved when training with 200 notebooks. We report on two metrics: Accuracy and Matthews correlation coefficient\(^7\). The best results are achieved by the Multi-Layered Perceptron classifier using the RDF2Vec embeddings method. With these results, we demonstrate that using a data node graph as an intermediate representation of the program improves the prediction accuracy systematically across all machine learning algorithms.

<table>
<thead>
<tr>
<th>Notebook Count</th>
<th>Average KG Size (Nodes)</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>8885</td>
</tr>
<tr>
<td>100</td>
<td>89202</td>
</tr>
<tr>
<td>200</td>
<td>186457</td>
</tr>
</tbody>
</table>

Table 4

The size of knowledge graphs per number of notebooks used across experiments.

6.4. Knowledge compression

The output of the machine learning application phase is a datanode graph whose nodes are all annotated with activity types. However, such representation does not allow us to see the intended activities at a higher level of abstraction. In this final phase, we apply the algorithm described in Listing 2 in order to generate a summarised view of the program by creating instances of activities in our ontology, absorbing adjacent data nodes annotated with the

\(^7\)https://scikit-learn.org/stable/modules/generated/sklearn.metrics.matthews_corrcoef.html

<table>
<thead>
<tr>
<th>Method</th>
<th>Accuracy</th>
<th>Matthews</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>mean std</td>
<td>mean std</td>
</tr>
<tr>
<td>CodeBERTa</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Decision Tree</td>
<td>0.49 0.36</td>
<td>0.32 0.48</td>
</tr>
<tr>
<td>GaussianNB</td>
<td>0.66 0.25</td>
<td>0.61 0.25</td>
</tr>
<tr>
<td>LinearSVC</td>
<td>0.43 0.45</td>
<td>0.25 0.58</td>
</tr>
<tr>
<td>LogisticRegression</td>
<td>0.05 0.00</td>
<td>0.00 0.00</td>
</tr>
<tr>
<td>MLPClassifier</td>
<td>0.40 0.40</td>
<td>0.30 0.45</td>
</tr>
<tr>
<td>RandomForestClassifier</td>
<td>0.52 0.31</td>
<td>0.33 0.45</td>
</tr>
</tbody>
</table>

RDF2Vec

<table>
<thead>
<tr>
<th>Method</th>
<th>Accuracy</th>
<th>Matthews</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>mean std</td>
<td>mean std</td>
</tr>
<tr>
<td>Decision Tree</td>
<td>0.68 0.06</td>
<td>0.58 0.07</td>
</tr>
<tr>
<td>GaussianNB</td>
<td>0.84 0.02</td>
<td>0.79 0.02</td>
</tr>
<tr>
<td>LinearSVC</td>
<td>0.95 0.01</td>
<td>0.93 0.02</td>
</tr>
<tr>
<td>LogisticRegression</td>
<td>0.95 0.01</td>
<td>0.93 0.02</td>
</tr>
<tr>
<td>MLPClassifier</td>
<td><strong>0.96</strong> 0.01</td>
<td><strong>0.95</strong> 0.01</td>
</tr>
<tr>
<td>RandomForestClassifier</td>
<td>0.85 0.01</td>
<td>0.80 0.02</td>
</tr>
</tbody>
</table>

Table 5

Results for test regime R1 for multiple classifiers using a KG based on 200 notebooks
same activity type. Here, we evaluate the ability of the approach to generate a more abstract representation of the process, by reducing the number of nodes required to express the data science pipeline. Specifically, we compare the size (number of arcs) of the data node layer with the one of the activity layer of the data journeys produced. Figures 2 displays the distribution of this compression factor, computed as the ration between the number of arcs of the data node graph and the ones of the activity graphs. The numbers demonstrate that our approach provides an effective way of summarising the data science pipelines, with almost all of the activity graphs being more than half of the size of the respective data node ones. Crucially, there has been a reduction of a 0.8 factor in about 3/4 of the cases.

Following the example introduced at the beginning of this section, the random-forests data journey contains 84 entities in the data node layer and only 18 in the activity graph layer, for a compression factor of 0.78. Figure 3 shows the activity graph (the equivalent data node graph is in Figure 1). The first representation accurately describes the data flow, but it is also very difficult to explore. The second, is a much more compact representation of core activities. Crucially, our layered approach allows us to use the more synthetic representation as a proxy to the underlying one and, indirectly, to the actual source code.

7. Conclusion

In this article, we proposed a rich ontology for representing Data Journeys. Using a corpus obtained from the 1000 most popular python notebooks on Kaggle, we showed how data journeys can be automatically identified from source code. The abstract representations that are generated are in many cases 80% the size of the original graphs. Additionally, our experiments show that an intermediate (datanode) graph representation (via RDF2Vec) improves performance over the use of just the code itself (via CodeBERTa) on the task of identifying high-level datascience activity. There are several limitations to our current approach. First, the current implementation is limited to Python
code, although the approach is portable to other languages. In addition, the implementation can be improved as there were notebooks that could not be properly processed. Here, we focused on the Activity Graph layer of the data journey. We plan to expand our approach to include also the other components of the Data Journey ontology, specifically, by decorating the datanode layer with richer types and relations specified in the Data Journey Ontology.

Importantly, we believe that data journeys are an exciting foundation for further work. In terms of use cases, data journeys, can provide a mechanism to more easily audit the data science models that are behind many web applications. They can provide a foundation for compliance checking in terms of correct licensing, data privacy and attribution. In terms of open science, data journeys can provide the possibility to wire in expertise and explanation at different levels of abstraction.

On a technical front, it would be interesting to develop methods that use data journeys to more effectively perform data and algorithm debugging. Likewise, developing more sophisticated abstraction models that can take advantage of both the code and graph representations is an exciting avenue of research.

We see a potential future in which data journeys with multiple layers of information about tasks, data, expertise, bias, can be used to build rich environments for transparency.

References
