Artificial Intelligence 302 (2022) 103627

Contents lists available at ScienceDirect

Artificial Intelligence

www.elsevier.com/locate/artint

Knowledge graphs as tools for explainable machine learning: A survey ☆

Ilaria Tiddi *, Stefan Schlobach

Vrije Universiteit Amsterdam, De Boelelaan 1111, 1081 HV Amsterdam, the Netherlands

ARTICLE INFO

Article history:
Received 1 May 2020
Received in revised form 12 October 2021
Accepted 18 October 2021
Available online 25 October 2021

Keywords:
Explainable systems
Knowledge graphs
Explanations
Symbolic AI
Subsymbolic AI
Neuro-symbolic integration
Explainable AI

ABSTRACT

This paper provides an extensive overview of the use of knowledge graphs in the context of Explainable Machine Learning. As of late, explainable AI has become a very active field of research by addressing the limitations of the latest machine learning solutions that often provide highly accurate, but hardly scrutable and interpretable decisions.
An increasing interest has also been shown in the integration of Knowledge Representation techniques in Machine Learning applications, mostly motivated by the complementary strengths and weaknesses that could lead to a new generation of hybrid intelligent systems. Following this idea, we hypothesise that knowledge graphs, which naturally provide domain background knowledge in a machine-readable format, could be integrated in Explainable Machine Learning approaches to help them provide more meaningful, insightful and trustworthy explanations.
Using a systematic literature review methodology we designed an analytical framework to explore the current landscape of Explainable Machine Learning. We focus particularly on the integration with structured knowledge at large scale, and use our framework to analyse a variety of Machine Learning domains, identifying the main characteristics of such knowledge-based, explainable systems from different perspectives. We then summarise the strengths of such hybrid systems, such as improved understandability, reactivity, and accuracy, as well as their limitations, e.g. in handling noise or extracting knowledge efficiently. We conclude by discussing a list of open challenges left for future research.

1. Introduction

The goal of this work is to study the integration and the role of knowledge graphs in the context of Explainable Machine Learning. Explanations have been the subject of study in a variety of fields for a long time [1], but are experiencing a new wave of popularity due to the recent advancements in Artificial Intelligence (AI) including machine and deep learning systems, which are now widely adopted for decision making. A major drawback, however, is their inability to explain their decisions in a way that humans can easily understand them, urging for the need to improve the interpretability and trustworthiness from the user perspective – a crucial aspect for their adoption at large scale. In response, Explainable Machine Learning has rapidly become an active area of research, with an explosion of contributions focusing on using

☆ This paper is part of the Special Issue on Explainable AI.

* Corresponding author.
E-mail addresses: i.tiddi@vu.nl (I. Tiddi), k.s.schlobach@vu.nl (S. Schlobach).

https://doi.org/10.1016/j.artint.2021.103627

a variety of techniques (e.g. visual cues, anchors, saliency maps or counterfactuals [2–4]) to elicit the decisions of both scrutable and inscrutable (black box) methods [5,6].

From a broader AI perspective, opaqueness is only one of the very well known limitations of modern subsymbolic systems – along with the need of large training data (data hunger), the poor ability to generalise across tasks (brittleness), lack of causal or analogical reasoning (reactivity) [7]. Interest has grown in the literature of the latest years, with the goal of fostering the integration of symbolic, knowledge-driven AI (i.e. Knowledge Representation) and subsymbolic, data-driven AI (Machine Learning), mostly motivated by the complementary strengths and weaknesses that could lead to designing hybrid, more intelligent systems [8]. The idea of neuro-symbolic AI is that symbolic methods, allowing to encode knowledge in the form of language-like, structured propositions that can be endlessly recombined to allow high-level reasoning across tasks and domains, could be combined with subsymbolic approaches and their ability to deal with large amounts of data, to handle noise, and to capture the richness of perceptual data. In this sense, it is natural to hypothesise that a neuro-symbolic integration could also support explainable systems to be more explainable, transparent and trustworthy.

In this work, we focus particularly on the role of knowledge graphs in the context of Explainable Machine Learning. Knowledge Representation has a long tradition in manipulating, creating, standardising, and publishing structured knowledge. In the last two decades, efforts have been focusing towards scaling up techniques to deal with the pervasive nature of the Web [9]. Semantic technologies allow to easily access knowledge sources on the Web, while symbolic representations in the form of ontologies, knowledge bases and graphs data-bases allow to formalise and capture knowledge and data about specific domains as well as general, encyclopaedic, knowledge. Such formal knowledge, which we will refer to as knowledge graphs, is machine readable, mostly publicly accessible and, more importantly, linked across domains – allowing machines to discover knowledge in structured but also serendipitous way [10,11]. The main goal of this paper is to investigate in which way knowledge graphs can be integrated in Explainable Machine Learning to provide more meaningful, insightful and trustworthy explanations.

To achieve this, we explore the landscape of Explainable Machine Learning in which subsymbolic systems have integrated structured knowledge at large scale, in order to identify the characteristics, strengths and limitations of such a hybrid integration. Using an approach based on a systematic literature review, we analyse different Machine Learning applications using structured knowledge in the form of graphs to generate explanations (we call these *knowledge-based explanations*), ranging from the earlier rule mining tasks, through classical tasks in image recognition, item recommendation, to natural language processing and predictive tasks.

This research is not intended as a survey of the whole field of eXplainable AI and Knowledge Representation, but has a particular focus on the advantages and limitations of using knowledge graphs as support and background knowledge for explainable systems. In particular, we present the following contributions:

- we present an analytical framework to systematically classify explainable, knowledge-based subsymbolic systems;
- we shape and organise the landscape of Knowledge-Based Explainable Machine Learning according to the above, particularly detecting strengths and limitations of each area;
- we discuss advantages and disadvantages of using knowledge graphs as background knowledge in the context of Explainable Machine Learning;
- ultimately, we provide the open challenges to be tackled in order to foster the design of next wave of explainable systems, integrating symbolic and subsymbolic reasoning in full.

2. Preliminaries

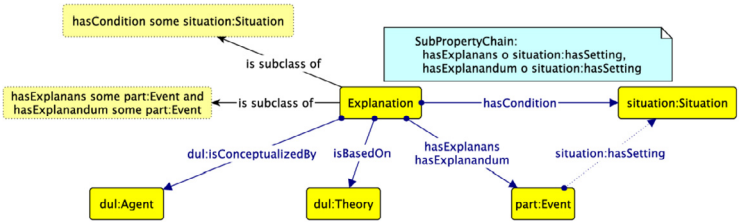
In order to lay the relevant ground for our analytical study, a preliminary step consists in establishing a working definition for explainability and provide the main notions regarding knowledge graphs. We achieve this by summarising the main theories around explanations with an historical overview before Machine Learning, and then providing a brief introduction on what are knowledge graph at scale and which techniques exists to manipulate them.

2.1. Overview of explanations before AI

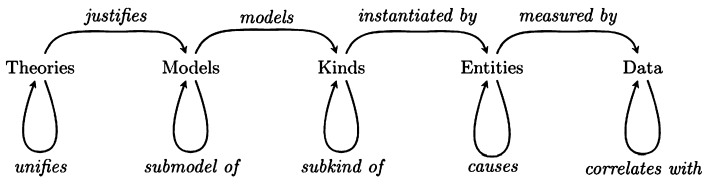
> *"The word explanation occurs so continually and holds so important a place in philosophy, that a little time spent in fixing the meaning of it will be profitably employed."* (John Stuart Mill, 1884)

Perhaps two centuries after Mill's wish, the success of precise but inscrutable models has pushed researchers from fields such as of cognitive science, law and social sciences to join forces with the Machine Learning community and work towards providing a unified view over the concept of explanation. Indeed, our view aligns with those studies arguing that explainable AI does need insights from those disciplines that extensively discussed explanations over time [1,12–14]. We refer the reader to the cited work for an in-depth discussion on the topic; here, we limit ourselves to identify a working definition of explainability for our research through highlighting how the different disciplines have perceived the concept of explanations across time.

Throughout history, philosophers have been looking at explanations as deductive or inductive situations where a set of initial elements (an event and some conditions) needed to be put into a relation with an consequent phenomenon accord-



(a) Explanation model as presented in [1].



(b) Explanation model as presented in [23].

Fig. 1. Examples of conceptual models of generic explanations as presented in recent literature.

ing to a set of empirical/metaphysical laws (see Aristotle's four causes of explanations,[1] Mill's scientific explanations [15], Hempel's deductive-nomological model [16]). Psychologists have been focusing on defining explanations as cognitive-social processes, see folks psychology [17], belief-desire-intention models,[2] script theory [18]), while linguists and focusing on explanations as the process of transferring knowledge (explanations as conversations [19], argumentation theories [20], Grice's maxims [21]. Finally, with the advent of Artificial Intelligence, explanations have been mostly seen processes where some initial facts and prior knowledge could be mapped to new knowledge under specific constraints (see rule-based models such as expert systems and inductive logics [22]).

While it is clear that a common agreement on a definition was never reached, the aspect that can be remarked here is that disciplines do share a common abstract model to define explanations, composed of law, events linked by causal relationships, and circumstances that constrain the events. Similar abstract models have been identified by recent work (cfr. [1] and [23] in Fig. 1). Following these, we will loosely refer to explanations as answers to why questions, in the form of "*when X happens, then, due to a given set of circumstances C, Y will occur because of a given law L*". Once identified a unified working model, we can look at how all explanation components (events, circumstances, law, causal relationships) can be structured as knowledge to be further used to generate explanations.

2.2. Structuring knowledge in the form of large-scale graphs

Although the idea of an intelligent model encoding real-world "things, not strings" and their (inter-)relationships was already present in the literature since the '80s [9], the term knowledge graph has regained popularity since Google announced the their Knowledge Graph in 2012.[3] Yet, there does not seem to be a precise definition for the term knowledge graph even to date [24]. The general agreement is to consider a knowledge graph as a data structure describing entities and their relationships by means of a directed, edge-labelled graph, often organising them in an ontological schema, and also covering several topics. One refers to the set of *concepts* and *properties* between them as Terminology Box (TBox) and to the set of *statements* about *individuals* belonging to those concepts as Assertion Box (ABox). With the increasing amount of techniques to scale up the Web, knowledge graphs are now result of the integration, extraction and manipulation of data from diverse sources at large scale. If following the semantic web standards,[4] one can refer to knowledge graph(s) as Linked Data.[5]

One of the main benefits of structuring knowledge in the form of graphs instead of typical relational settings is the flexibility towards the schema, that maintainers can define at a later stage, and change over time. This allows more flexibility for data evolution, as well as capturing of incomplete knowledge [9]. Reasoning over knowledge graphs can be performed by means of standard knowledge representation formalisms (RDF, RDFS, OWL), allowing to describe and label entities and

[1] https://plato.stanford.edu/entries/aristotle-causality/#FouCau.
[2] https://plato.stanford.edu/entries/folkpsych-theory/.
[3] https://www.blog.google/products/search/introducing-knowledge-graph-things-not/.
[4] https://www.w3.org/wiki/LinkedData.
[5] https://www.w3.org/standards/semanticweb/data.

Table 1
Summary of most common KGs in the literature, adapted and updated from [28].

Name	Domain	TBox		ABox		Version
		Concepts	Properties	Individuals	Statements	
OpenCyc	Common-sense	45k	19k	240k	2M	2012
Freebase	Factual	27k	38k	50M	3B	2015
Wikidata	Factual	23k	1,6k	1.5M	714M	2019
DBpedia	Factual	754	2.9k	5,1M	400M	2016
Yago3	Factual	488k	77	17M	1.2B	2017
ConceptNet	Common-sense	200k	36	32M	21B	2019
WordNet3.1RDF	Domain	4	6	155k	2.6M	2014

the relationships between them. Query languages such as SPARQL,[6] Cypher,[7] Gremlin,[8] allow both standard relational operations and navigational operators that allow to find arbitrary-length paths between entities, supporting a more advanced knowledge discovery. Additionally, more advanced graph manipulation such as analytics, summarisation, completion can be performed through frameworks able to deal with knowledge graphs at scale [25–27].

A number of knowledge graphs have been made available on the Web in the last years also thanks to a variety of standards and practices for data representation, publishing and exchange [28]. The most adopted KGs in the literature are presented below and summarised in Table 1 along with some statistics. We additionally categorised them according to three categories, i.e. "common-sense KGs" that contain knowledge about the everyday world, "factual KGs" containing knowledge about facts and events, and "domain KGs" that encode knowledge from a specific area (linguistics, biomedical, geographical etc.).

- *OpenCyc*,[9] a curated knowledge graph of common-sense knowledge implemented from the Cyc knowledge base;
- *Freebase*,[10] a KG built as a crowdsourced effort using structured data from a number sources including Wikipedia schemes and contributions;
- *WikiData*,[11] also a free, collaborative project build on top of wiki contents, additionally providing metadata about data provenance;
- *DBpedia*,[12] a knowledge graph built by automatically extracting pairs of key-values from the Wikipedia infoboxes, which are then mapped to the DBpedia ontology with crowdsourcing;
- *YAGO*,[13] a large KG which maps facts from WikiData, GeoNames and other data sources to a taxonomy build by combining WordNet and Wikipedia categories;
- *ConceptNet*,[14] a free, crowdsourced linguistic KG including information from WordNet, Wikipedia, DBpedia and OpenCyc
- *WordNet*,[15] a curated lexical database of relationships between concepts (hypernyms, pertainyms, meronyms etc.). We refer to its RDF version of 2014, WordNet3.1RDF.

Note that some of the above resources have been discontinued, while a number of other freely available sources exists (e.g. NELL [29], KBpedia,[16] FrameNet[17]) whose usage has been so far limited in the literature. Moreover, a number of proprietary KGs, sometimes called Enterprise Knowledge Graphs (EKGs) have been created in the latest years – see Google's and Facebook's Knowledge Graph, Amazon's and Ebay's Product Graphs.[18] Finally, we also consider KGs domain ontologies without assertions, e.g. schema.org,[19] that might have been ad-hoc built in order to support specific tasks.

2.3. Explainable AI needs structured knowledge: practical examples

The need of modern Machine Learning methods to be more transparent in their decisions has become evident after events as the Cambridge Analytica scandal and the disruptions of the 2016 US elections.[20] A number of initiatives have

[6] https://www.w3.org/TR/rdf-sparql-query/.
[7] https://neo4j.com/developer/cypher-query-language/.
[8] https://tinkerpop.apache.org/gremlin.html.
[9] http://www.cyc.com/opencyc/a, discontinued in 2017.
[10] http://www.freebase.com, discontinued in 2015.
[11] https://www.wikidata.org/wiki/Wikidata:Main_Page.
[12] https://www.dbpedia.org.
[13] https://github.com/yago-naga/yago3.
[14] http://conceptnet.io/.
[15] https://wordnet.princeton.edu.
[16] http://kbpedia.org/.
[17] https://framenet.icsi.berkeley.edu/fndrupal/.
[18] For numbers and a detailed description, see [30].
[19] https://schema.org/.
[20] https://www.theguardian.com/technology/2017/may/22/social-media-election-facebook-filter-bubbles.

been launched ever since, e.g. DARPA's eXplainable AI program [31] and the EU Ethical guidelines for Trustworthy AI[21] launched to encourage the design ethical systems that humans would appropriately understand, manage and trust.

Most of the Explainable Machine Learning methods, however, still focus on the interpretation of the inner functioning of black-box models, such as identifying the input features that are associated the most with different outputs using visual cues, anchors, saliency maps or counterfactuals [2–4]. While this represents a mere approximation of the complex decision-making functions, the models are neither able to account for any context yet nor for background knowledge that users might possess [32]. Leaving such methods to take decisions for life-critical problems such as healthcare is then problematic, calling for a new paradigm of intelligent systems that deliver understandable results along with transparent, reliable explanations. Knowledge graphs and the structured Web represent a valuable form of domain-specific, machine-readable knowledge, and the connected or centralised datasets available could serve as background knowledge for the AI system to better explain its decisions to its users. Below we present a few scenarios for the medical domain, where the Machine Learning system could benefit from external knowledge to support domain experts in understanding why the algorithms came up with certain results.

Counterintuitive predictions The work of [33] reports of a model counterintuitively predicting that asthmatic patients have a lower risk of dying from pneumonia. In order to explain such decisions, the doctor's medical expertise is required to reveal that these patients were admitted directly to the Intensive Care Unit, receiving an aggressive care that indeed lowered their risk of death, but also caused incorrect machine-driven conclusions. Such decisions could have been more understandable if the model was also providing evidence in the form of explanations found in the external, machine-readable knowledge sources, available, e.g., hospital databases providing patients' history or drug-disease interaction datasets such as DisGeNET.[22]

Clinical trial recommendation Let us imagine a health care company that uses an AI system to supports cancer-diagnosed patients in finding experimental treatments (early access programs or EAPs). A patient provides the system with a description of his medical history (relevant documents, symptoms, diagnosis, etc.), which in turn extracts the salient information using text analysis, then finds the list of candidate clinical trials from the company's knowledge base using the search engine component. A medical expert in the company then has to verify the identified trials based on his expertise, and present a full report to the patient. The expert's time and efforts could be considerably reduced when combining the patient's data with structured knowledge such as medical ontologies, thesauri, PubMed's evidence about previous studies, and provide the expert not only with a list of relevant clinical trials, but also with explanations of why these were selected. This process would guarantee that (a) the expert's knowledge is not substituted, but rather complemented and integrated in the overall process, and (b) trust is increased by showing that the results were obtained using reliable domain expertise augmented by an automated system.

Treatment diagnosis The work of [34] shows how a user-centred AI system for diagnosis recommendation requires clinicians to complement the intelligent agent with their own explanations about the patient's case. A user-study is carried out to identify the different types of explanations required at the different steps of the automated reasoning, i.e. "everyday explanations" for diagnosis, "trace-based explanations" for planning the treatment, "scientific explanations" to provide scientific evidence from existing studies, and "counterfactual explanations" to allow clinicians to add/edit information to view a change in the recommendation. Ontologies are used to model the components (primitives) necessary for the AI system to automatically compose explanations exposing different forms of knowledge and address different tasks performed by the agent.

3. Research methodology

Our research methodology is mostly articulated in the common three steps of qualitative research, namely the identification of an analytical framework comprehending the main variables to analyse, the assessment of such variables through literature exploration, and literature synthesis and discussion.

3.1. Research questions and analytical framework

As stated in Section 1, our main research question is *how can the large-scale knowledge graphs be integrated in Explainable Machine Learning systems to provide more trustworthy explanations*? To answer this, we focus on analysing the literature following the subquestions below:

1. Which **characteristics** have knowledge graphs that are employed by a subsymbolic system generating knowledge-based explanations? Which type of knowledge they represent (domain, factual, common-sense knowledge), how expressive they are (ABox, TBox, both)? Was the knowledge to generate explanations extracted automatically (e.g. by following links in

[21] https://ec.europa.eu/digital-single-market/en/news/commission-appoints-expert-group-ai-and-launches-european-ai-alliance.
[22] https://www.disgenet.org/.

Table 2
Analytical toolkit to classify knowledge-based explainable systems. The (*) indicates a non-exhaustive list.

Variables	Name	Possible values
Knowledge (KG)	Type	domain (DK), factual (FK), common-sense (CK)
	Semantics	TBox (T), ABox (A), both (T/A)
	Selection	manual (man), automatic (aut)
	Reusability	✓, ✗
	Number of Knowledge Graphs	integer
Model (ML)	Input*	tabular data (tab), images (img), text (txt), other (o)
	Method*	rule- (r), tree-based (t), neural nets (nn), other (o)
	Task	classification (cls), prediction (prd), regression (reg)
	Integration	internal (int), external (ext) knowledge integration
Explanation (XP)	Form*	raw data (raw), written language (nl), visual (viz), multimodal (mm)
	Type	functional, categorical (cat), mechanistic (mec), functional (fct)
	Interpretability	Post-hoc (pos) or integrated (int), or hybrid (hyb)

the graph) or manually (e.g. using an expert to select relevant portions of the graph)? And were explanations built by reusing existing knowledge graphs, or ad-hoc built structures? How many graphs were taken into consideration? All these questions are grouped as the *Knowledge* (KG) variables.

2. Which **type of subsymbolic approaches** are able to generate knowledge-based explanations? And particular, which type of input data does the model handle (tabular data, images, textual data), which task is the model used for, and which method is used to (sequential neural networks, convolutional, tree-based etc.)? Also, how is the model integrating structured knowledge internally (i.e. are statements embedded in the global system behaviour) or externally (i.e. used a posteriori)? We call these aspects *Model* (ML) variables.
3. Which **type of explanations** is the system dealing with? In which form are they communicated (text/natural language, visual images etc.), are explanations categorical (explaining the properties of a result), mechanistic (the mechanisms causing a result) or functional (explaining the behaviour and end-goal of something)[23]? Finally, does the explanation content concerns the output of the model, its behaviour, or a combination of the two (post-hoc or integrated interpretability, or hybrid)? These are defined as *Explanation* (XP) variables.

All these variable can be organised in the analytical toolkit shown in Table 2. In the next sections, we will use this toolkit to first analyse and discuss a set of relevant studies, and to finally to identify a few guidelines to improve on the integration of knowledge graphs in Explainable Machine Learning.

3.2. Literature search and selection

All studies were searched in one of the two following ways: a comprehensive top-down approach to extensively search Knowledge Representation papers from the major academic databases, IEEExplore, ACM Digital Library, Google Scholar and ScienceDirect. This was accompanied by a bottom-up approach to check workshops and Artificial Intelligence conferences for freshly published research outputs in the area of explainable AI, additionally snowballing through the cited literature.

We keyword-searched for works containing a conjunction of any of the terms summarised in Table 3, leading to a selection of more than 10,000 articles. Approx. 150 were thoroughly scanned according to the criteria that follow, and about 100 more were identified directly from the related work sections. Whenever possible, we prioritised peer-reviewed publications and major journals/conferences to white papers or unreviewed submissions. Studies were selected only if presenting a subsymbolic system including some form of learning from data, and producing any kind of explanation using background knowledge in the form of graphs. This means, for instance, that expert systems, case-based reasoning, and other rule-based approaches are not in the scope of this work, along with planning and decision making applications. Through this approach, we finally identified a set of 53 papers.

Both the analytical toolkit and the studies presented and analysed in the next sections are openly available as part of the Open Research Knowledge Graph [35].[24] This contributes to future research on the same topic by enabling consultation of existing literature in a structured manner and make relevant comparisons (cfr. snapshot in Fig. 2). It will also promote the analytical toolkit in the form of vocabulary for reuse.

4. Using knowledge graphs for explainable machine learning

The AI literature has a long history of working on exploiting structured knowledge for a system's explainability, starting already from the early knowledge-based systems designed in response of a first generation of expert systems unable

[23] Also called respectively factual, scientific and behavioural explanations, these categories correspond to Aristotle's 4 modes of explanations. *Material explanations* are excluded as considered irrelevant to this work. See [12] for further discussions.
[24] Comparison at https://www.orkg.org/orkg/comparison/R69680, DOI: https://doi.org/10.48366/r69680.

Table 3
Summary of the keywords used in the literature search.

Area	Keywords
Explainability	explanations, interpretation, explainability, explication, elicitation, interpretability
Subsymbolic AI	black box, inscrutable models, deep learning, machine learning, neural networks, classification, regression, prediction, subsymbolic reasoning, subsymbolic AI, statistical AI
Symbolic AI	knowledge graphs, ontologies, structured knowledge, knowledge bases, Linked Data, knowledge-based systems, symbolic reasoning, symbolic AI

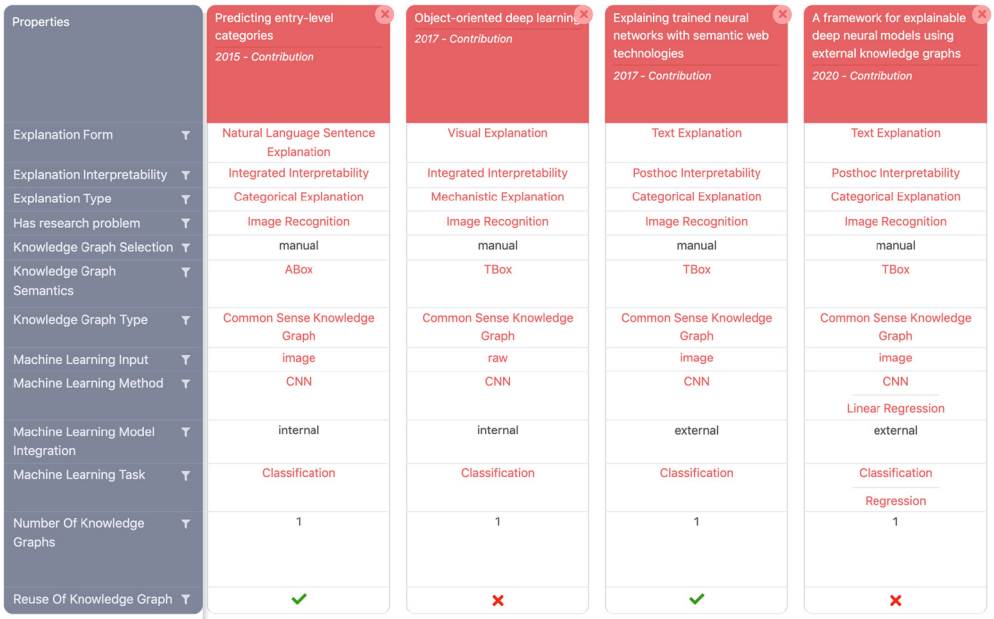


Fig. 2. Comparison of the identified literature in a structured format, as part of the ORKG initiative.

to provide trustworthy justifications [36,37]. We focus here on Machine Learning-based applications and organised works according to their generic application domain – from rule-mining approaches, to image classification and item recommendation tasks, to natural language applications and predictive tasks. The organisation is purely pragmatical and made to maximise the logics in our narrative; yet, we tried to maintain an historical order from earlier, simpler approaches to more complex, modern architectures.

4.1. Rule-based machine learning

Knowledge Discovery became an established field towards the end of the '80s and saw a large body of work focusing on explaining the outputs of rule-based Machine Learning algorithms able to identify interesting patterns in large amounts of data (the so-called "data post-processing" or interpretation step in the KD pipeline [38]).

Structured knowledge in the form of domain ontologies was investigated with the idea that it could to support (or potentially replace) experts in this data interpretation step – cfr. seminal work of [39] to translate the outputs of a neural network into symbolic knowledge using a domain ontology in the form of Horn clauses. This idea was then further expanded by [40,41] to explain generic data mining patterns with hand-crafted ontologies. This resulted particular useful to analyse or filter association rules in the biomedical domain, be the background knowledge in the form of existing taxonomies [42], meta-thesauri [43] or hand-crafted domain ontologies [44]. Similarly, [45,46] use domain knowledge the Gene Ontology and the KEGG ontology for subgroup discovery, with the idea that the constructed rules describing subgroups are good explanations for their forming. With the rise of Linked Data, several approaches suggested the use of links across datasets explain sequence patterns [47–49] through graph exploration. This idea was also explored to explain data to a non-expert audience, especially through augmenting tabular data and statistical analyses using multiple datasets as DBpedia, Eurostat, and GADM [50–52].

Table 4
Knowledge-based explanations in early Machine Learning.

	Knowledge graphs					Model				Explanations		
	Type	*Semantics*	*Selection*	*Reusability*	*Number of KG*	*Input*	*Method*	*Task*	*Integration*	*Form*	*Type*	*Interpretability*
[39]	DK	T	man	✗	1	raw	nnet	cla	ext	txt	mec	pos
[41]	FK	A	man	✗	1+	mm	rb	clu	ext	o	cat	pos
[40]	DK	A	man	✗	1	mm	rb	clu	ext	txt	mec	pos
[42]	DK	T	man	✓	1	raw	rb	clu	ext	txt	mec	pos
[44]	DK	T	man	✗	2	raw	rb	clu	ext	txt	mec	int
[43]	DK	A	man	✓	2	raw	rb	clu	int	txt	mec	pos
[45]	DK	T	man	✓	1	raw	rb	clu	int	txt	cat	pos
[46]	DK	T	man	✓	1	raw	rb	clu	int	txt	cat	pos
[47]	DK	A	man	✓	1	raw	rb	clu	int	txt	mec	pos
[48]	DK	A	man	✓	1	raw	rb	clu	ext	txt	mec	pos
[49]	FK	A	man	✓	1	raw	rb	clu	ext	txt	mec	pos
[50]	FK	A	man	✓	1	tab	o	reg	ext	img	cat	pos
[51]	FK	A	man	✓	3	tab	o	reg	ext	img	cat	pos
[52]	FK	A	man	✓	1	tab	o	reg	ext	img	cat	pos
[53]	FK	A	aut	✓	1+	raw	ilp	clu	ext	nl	cat	pos
[54]	FK	A	aut	✗	1	rb	clu	-	int	txt	mec	int

Background knowledge from several Linked Data sources is also used by Dedalo [53], where an inductive logic-based graph search is performed to build path-based explanations of data output by unsupervised learning algorithms (clusters, association rules, time series). These approaches exploit the "follow your nose principle" of publishing and linked information on the Web, hence avoiding the a priori selection of specific knowledge graphs to reason upon. This allows to derive explanations with information gathered through a deeper graph exploration (i.e. not restricting to the nodes' closest neighbourhood), but comes at costs of making no use of logical inference, and building atomic explanations from ABox statements only. ILP also inspired the neural forward-chaining differentiable rule induction network of [54]. TBox and ABox statements are represented using dense vectors, which are trained using gradient descent aimed at learning the best representations for a given predicate. To cope with the computational costs of reasoning, the authors use an ad-hoc taxonomy of `is-a`, `has-a` relationships.

A main feature of these systems, summarised in Table 4, is that the structured knowledge to generate explanations is integrated a posteriori, i.e. once obtained the outputs of the models. With a few exceptions, these works are limited in what they rely on the manual selection of the knowledge graphs, requiring an expert to extract from these useful background knowledge in the form of statements. When relying reusing existing knowledge graphs to avoid the time-consuming knowledge acquisition steps, their performance is highly dependent on the freshness of the information stated in ABox assertions, by nature more dynamic and subject to expiration over time. The use of domain knowledge graphs which turned with the use of factual knowledge afterwards, alongside with explanations becoming more visual and fact-based, suggests that these systems moved from targeting an audience of domain-experts that could understand articulated explanations, to one where users would need visual support to better understand their decision and, consequently, trust them. Examples of such explanation types, where properties and values from Linked Data are used to interpret observations, can be seen in Fig. 3.

4.2. Image recognition

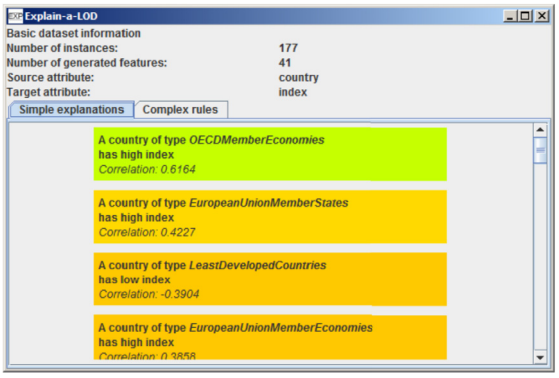
Early work exploiting structured knowledge for Machine Learning-based visual explanations was focused on the task of image recognition, e.g. in [55] a manually-curated ontology of spatial concepts, colours, textures and their relationships is incorporated in a multi-layer perceptron classifier and used to identify Brodatz texture in images. The main insight here is the use of the domain knowledge to foster the transparency in the process, acting as a user-friendly intermediate between the classifier and the end-user.

This line of work continues to this day, where large-scale, publicly available knowledge graphs are embedded in more scalable learning algorithms to visually explain the model behaviour (e.g. visualising hidden states of a neural network). For example, [56] show how background knowledge from ConceptNet can be exploited for explaining objects in images in the form of captions, through incorporating into a set of visual detectors both concepts (e.g. *Dish*, *Person*, *Kitchen*) and relationships (e.g. *atLocation*, *UsedFor*, *MadeOf*) connected to a image retrieval keyword (e.g. *Chef*). Successful experiments ran on Microsoft's COCO dataset[25] demonstrate that the integration of knowledge graphs is a valuable research line to be

[25] http://cocodataset.org/.

explain($\mathcal{Y}$): countries where males are more educated		
exp_i	F(%)	Time"
⟨*skos:exactMatch, dbp:hdiRank.*≥.“126”⟩	87.8	197”
⟨*skos:exactMatch, dc:subject.* **db:Category:Least_developed_countries**⟩	74.7	524”
⟨*skos:exactMatch, dbp:gdpPppPerCapitaRank.*≥.“89”⟩	68.3	269”
⟨*skos:exactMatch, dc:subject skos:broader.* **db:Category:Countries_in_Africa**⟩	67.1	540”
⟨*skos:exactMatch, dbp:populationEstimateRank.*“76”⟩	61.9	201”
⟨*skos:exactMatch, dbp:gdpPppRank.*≥.“10”⟩	59.1	235”

(a) Dedalo [53].



(b) Explain-a-LOD [50].

Fig. 3. Knowledge-based explanations in early Machine Learning, using properties and values from Linked Data to explain observations.

explored, but reveal the need of focusing on filtering the data therein represented to identify relevant knowledge. Visual detectors integrated in a Convolutional Neural Network are also used by [57], that exploits Wordnet's fine-grained categories (assumed to be closer to what people are likely to name objects) to automatically predict object categories in an image. Experiments show that the model emulates the naming choices of human observers in a more effect way.

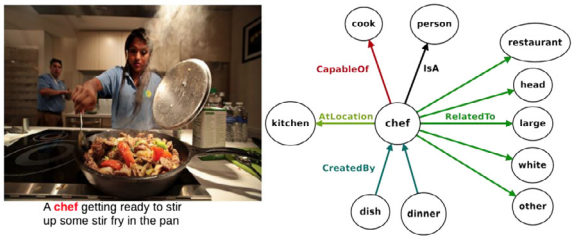
With the advent of complex deep learning architectures the idea of a neuro-symbolic integration with knowledge graphs has appeared also in image recognition tasks. An approach called Object-Oriented Deep Learning is presented by [58], where the N-dimensional tensor of the deep net architecture is replaced by object common-sense knowledge (with minimal properties such as position, pose, scale) with the goal of obtaining a more interpretable visual knowledge throughout all network layers. Multiclass prediction using non-propositional rules from a knowledge graph is presented in [59] in the context of object detection. To do so, a CNN architecture is combined to an ILP framework that allows to learn OWL class expressions based on background knowledge in the form of ontological rules. Similarly, authors of [60] combine a generic DNN architecture with WordNet for the task of scene classification. Object types from WordNet's are aligned to objects in the ADE20K dataset, and then use WordNet's hierarchy to train an object recognition module that is further fed into a linear regression model able to provide human-understandable explanations automatically. Both approaches rely on is-a relationships only, limiting the potential of using the graphs. This proof-of-concept paper also relies on classes of the Suggested Upper Merged Ontology (SUMO[26]) and one simple relation (*contains*(image,class)).

Insights in this area come from works suggesting that knowledge graphs valuably tackle the limitations of state-of-the-art models, e.g. task design efforts, inaccuracies, computational costs or data hunger. For example, [61] reports the use of knowledge graphs in a posteriori process to explain satellite images misclassified by a convolutional autoencoder. An existing geographical knowledge graph is first used to identify the images' structure (in terms of concepts and spatial relationships in their vicinity), and then to reason upon spatial constraints that could justify the errors made by the classifier. Authors of [62] integrate WordNet and the Visual Genome taxonomy in a graph search-based CNN that reasons about relationships and concepts in a given image, and produces outputs on the nodes representing visual concepts (used to classify objects in the image) for multi-label image recognition. Explanations on the classifications are derived by following the propagation of the information in the graph. [63] exploits a custom knowledge graph automatically built from images and existing meta-data sources integrated in a Markov Logic Network for zero-shot affordance prediction and object recognition avoiding the training of separate classifiers (for object labelling, attribute recognition, affordance detection). A Graph Convolutional Network that integrates both word embeddings and explicit knowledge from NELL to learn visual classification of unseen classes (zero-shot learning) is used by [64]. The idea of integrating both word-level knowledge and common-sense semantics expressed in knowledge graphs within deep network structures is also investigated by [65], that exploits WordNet to create

[26] http://www.adampease.org/OP/.

Table 5
Knowledge-based explanations for image recognition.

	Knowledge graphs					Model				Explanations		
	Type	*Semantics*	*Selection*	*Reusability*	*Number of KG*	*Input*	*Method*	*Task*	*Integration*	*Form*	*Type*	*Interpretability*
[55]	CK	A	man	✗	1	img	mlp	cls	int	img	cat	int
[56]	CK	A	man	✓	1	img	me	reg	int	nl	cat	int
[57]	CK	A	man	✓	1	img	cnn	cls	int	nl	cat	int
[58]	CK	T	man	✗	1	raw	cnn	cls	int	img	mec	int
[59]	CK	T	man	✓	1	img	cnn	cls	ext	txt	cat	pos
[60]	CK	T	man	✗	1	img	cnn+lr	cls+reg	ext	txt	cat	pos
[61]	CK	A	man	✓	1	img	cnn	cls	ext	txt	cat	int
[62]	CK	A	man	✓	1	img	cnn	cls	ext	img	cat	int
[63]	CK	A	man	✓	1	img	mln	cls	int	img	mec	int
[64]	CK	A	man	✓	1	raw	gcn	cls	ext	img	cat	int
[65]	FK	T	man	✓	3	mm	cnn	cls	ext	txt	mec	int



(a) MS Coco image including the word *Chef*, explained with the ConceptNet graph [56].

∃contains.Window	(1)	∃contains.LandTransitway	(6)
∃contains.Transitway	(2)	∃contains.LandArea	(7)
∃contains.SelfConnectedObject	(3)	∃contains.Building	(8)
∃contains.Roadway	(4)	∀contains.¬Floor	(9)
∃contains.Road	(5)	∀contains.¬Ceiling	(10)

(b) Explaining the concept *Warehouse* with relationships from the SUMO ontology [59].

Fig. 4. Knowledge-based explanations in image recognition tasks, where semantic restrictions are used to elicit information about images.

mappings between ImageNet and Wikidata. A pre-trained CNN is used to classify images captured using OpenCV, converting the outputs in WordNet synsets and retrieving information from the Wikidata item they are linked to.

These works, summarised in Table 5, present a number of novel features: the integration of knowledge graphs within the model, the use of graphs to explain (and adjust) how the model reached the conclusions, the use of common-sense knowledge graphs rather than factual ones (likely due to common-sense knowledge being represented in images), and the use of a variety of different models for generating knowledge-based explanations, showing the ability to knowledge graphs to generalise across tasks. A few examples are given in Fig. 4, where we can see how semantic restrictions are used to explain image outputs of the model. We note that the approaches are mainly focused on entailment relationships (i.e. subclasses), limiting the potential of knowledge graphs, and are characterised by the manual step of extracting/aligning knowledge sources, making emerge issues such as missing/wrong statements and entity ambiguity in knowledge graphs.

Table 6
Knowledge-based explanations for recommender systems.

	Knowledge graphs					Model				Explanations		
	Type	*Semantics*	*Selection*	*Reusability*	*Number of KG*	*Input*	*Method*	*Task*	*Integration*	*Form*	*Type*	*Interpretability*
[66]	FK	A	man	✗	1	tab	CNN	reg	int	txt	mec	int
[67]	FK	A	man	✗	1	tab	CF	cls	int	txt	mec	int
[68]	FK	A	man	✓	1	tab	rnn	cls	int	txt	mec	int
[70]	FK	A	man	✓	1	tab	rnn	cls	int	txt	mec	int
[71]	FK	A	man	✓	1	tab	rank	clu	ext	txt	cat	pos
[72]	FK	A	man	✓	3	tab	rank	clu	ext	nl	cat	pos

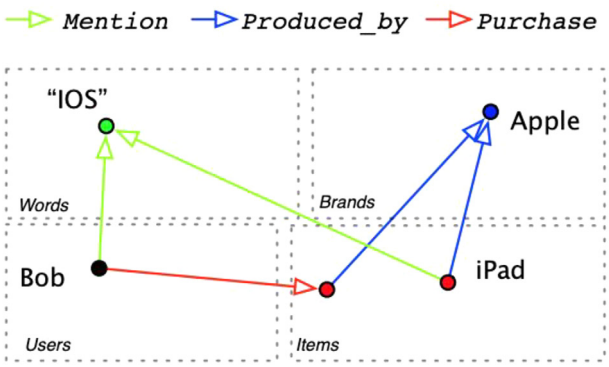
4.3. Recommender systems

Knowledge graphs to provide more transparent results to models' outputs have recently experienced a take-up also in the area of recommender systems, with the goal of enhancing the users' experience in terms of satisfaction, trust, and loyalty. Most of the approaches are content-based, i.e. they consists of explaining a recommendation with entities from a given knowledge graph in the form of images or natural language sentences.

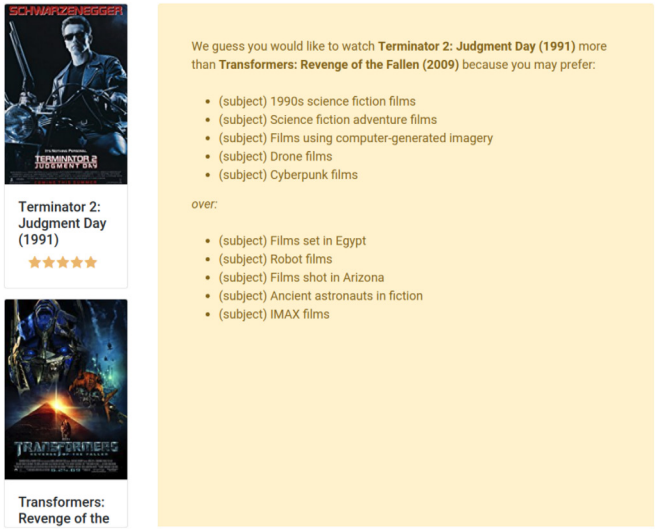
We find here mostly integrated approaches as, for instance, [66] proposing DKN as a deep neural network that incorporates a knowledge graph into a news recommender system. The authors tackle the task of click-through rate prediction, taking one piece of candidate news and the user's click history as input, giving the probability of the user clicking the news as output. Words in the news are automatically associated with entities (and their immediate neighbours) in the graph and finally embedded in a vector used by a CNN that predicts the user's clicking likelihood. An attention-based layer is also incorporated, in order to automatically match the user's history with candidate interesting news. Similarly, [67] focuses on recommending Amazon products by combining a collaborative filtering approach with a knowledge graph. An ad-hoc, automatically-built graph of entities and user behaviours, and a set of minimal properties (e.g. *produced by*, *category*, *also viewed*) is embedded in the model. The generated personalised recommendations and their explanations expressed in natural language are then built using a soft-matching algorithm over the knowledge graph. Along the same lines but reusing existing knowledge sources, authors of [68] propose to substitute the hidden layers and the connections of an autoencoder neural network with the structure of DBpedia, thus following the principles of the Semantics-Aware Autoencoders [69]. To define movies, authors use a small set of predicates (`dct:subject`, `dct:starring`, `dct:director`, `dct:writer`), and then formulate human-understandable explanations by relying on the weights associated to features in the user's profile. Explanations are presented in 3 forms, based on popularity ("*We suggest X and Y since they are very popular among people who like the same movies as you*"), pointwise personalisation ("*We guess you would like to watch something since they are about X and Y*") or pairwise personalisation ("*We guess you would like to watch X more than Y because you may prefer* x_i"). A very similar approach is the one of [70], that uses Freebase to enhance a sequential recommender system for movies and books that integrates a Recurrent Neural Network and a Key-Value Memory Network.

A post-hoc approach suggesting that knowledge graphs can be used not only to generate human-understandable explanations, but also to augment the model with additional knowledge on the task at hand is presented in [71] for generating natural language explanations to recommend movies using DBpedia on top of the ExpLOD framework. A ranking algorithm is proposed to rank the most relevant properties to items that a user likes, which are then used to build natural language explanations. Another interesting argument is brought forward by [72], arguing that explainable recommender systems are limited in what they only incorporate knowledge from the node's closest neighbours, and lack of proper solutions to filter unrelated entities. The authors propose to explain travel recommendations using the unstructured data of items that have rich textual content that can be valuable to build explanations (e.g. books, news, touristic tours...). DBpedia is used to filter out irrelevant entities (through the DBpedia categories), while a large-scale knowledge graph integrating DBpedia, schema.org and YAGO is used to augment information about entities and subsequently build explanations for the recommendations in natural language. See Table 6 for an additional summary.

The literature in explainable recommender systems here features aspects similar to the systems presented in the previous section, e.g. mostly models integrating knowledge graphs internally and explanations in written or visual form. Significant differences lie mostly in the use of a large-scale, open-domain knowledge graph (DBpedia, mainly) to extract additional facts about the input data that can explain the recommendation in a user-friendly way, and in the lack of use of terminology axioms, suggesting the difficulty in scaling up inference techniques [73]. Open-domain knowledge graphs also introduce issues such as information overflow, i.e. the high outdegree of the nodes limits the selection of the closest facts in the graph, often shown in the forms of multi-edge paths extracted from the graphs (cfr. examples of Fig. 5). This not only prevents a more extended knowledge discovery process, but often also requires an accurate pruning, in order to obtain explanations that end-users can better trust.



(a) Explanation paths between a user *Bob* and the recommended item *iPod* [67].



(b) Pairwise explanations of recommended movies using *dc:subject* relationships [68].

Fig. 5. Knowledge-based explanations for recommender systems in the form of multi-edge paths extracted from the graph.

4.4. Natural language applications

Taking inspiration from the social sciences arguing that explaining also involve a social process of communicating information from an *explainer* to an *explainee* [74], a body of relevant work can be identified in natural language applications such as knowledge-based Question-Answering (KB-QA), machine reading comprehension and Conversational AI in general, where knowledge graphs have been used mostly as background knowledge to answer common-sense knowledge questions both in the form of images, speech and text.

For example, [75] integrate both ConceptNet and WordNet in the Knowledgeable Reader, an attention-based model for reading comprehension, inferring the answer from a given document using external information retrieved from these sources. Authors of [76] combine ontological knowledge and reasoning capabilities of existing lexical knowledge bases (BabelNet, NASARI, and ConceptNet) in `MeRaLi`, which allows to build explanations accompanying scores in the task of conceptual similarity. The similarity between two terms is computed through quantifying the amount of information shared between the respective entity vectors in the COVER space. Explanations for the scores are generated in natural language by extracting the (explicit and human-readable) matching properties and values in two compared vectors. The quality of the automatically generated explanations is then assessed in a user-study. Similarly, [77] explains the semantic relationships holding between textual documents using a Distributional Semantic Model and knowledge from WordNet, in the form of natural language human-like justifications.

ConceptNet is also used as background knowledge in KB-QA to answer domain-specific questions, e.g. by combining query reformulation, structured background knowledge and textual entailment to explain answers to scientific questions [78] or providing commonsense links between concepts resulting from QA models [79]. Open-domain knowledge graphs are in-

Table 7
Knowledge-based explanations in Natural Language applications.

	Knowledge graphs					Model				Explanations		
	Type	*Semantics*	*Selection*	*Reusability*	*Number of KG*	*Input*	*Method*	*Task*	*Integration*	*Form*	*Type*	*Interpretability*
[75]	CK	T	man	✓	2	txt	mrc	cls	int	nl	mec	int
[76]	DK	T	man	✓	3+	raw	rb	cls	int	txt	mec	int
[77]	DK	T	man	✓	1	raw	tb	cls	ext	nl	mec	int
[78]	DK	T	man	✓	1	txt	dtr	cls	int	txt	cat	int
[79]	CK	T	man	✓	1	txt	gcn	cls	ext	img	cat	int
[80]	FK	A	man	✓	2	txt	rnn	cls	ext	txt	cat	int
[81]	FK	A	man	✓	3+	txt	bn	cls	ext	txt	cat	int
[82]	CK	T/A	man	✓	3	img	rnn	cls	ext	img	cat	int
[83]	CK	T/A	man	✓	1	img	rnn	cls	ext	img	cat	int
[84]	CK	T	man	✓	1	img	gcn	cls	int	txt	cat	int
[85]	FK	T	aut	✗	1	img	lp	cls	int	nl	mec	pos
[86]	CK	T	man	✓	3	txt	rb	cls	int	nl	mec	int
[87]	FK	T	man	✓	1	nl	-	cls	int	nl	mec	int
[88]	FK	A	man	✓	1	txt	-	-	ext	nl	mec	int

stead used by [80] to answer single-fact questions converted into structured SPARQL queries. A recurrent neural network (RNN) structured as Gated Recurrent Unit is used to produce the representation of questions, i.e. to detect the subject and relation mentioned in the question, further constructed as SPARQL queries to retrieve the desired entity as an object. The method is trained on the SimpleQuestions dataset consisting of approx. 110k English questions paired with a subject-relation-object triple from Freebase. The work of [81] is an attempt to extend this idea and use multiple, interconnected knowledge graphs to improve the open-domain question-answering. Candidate triple patterns (subject, property, object) are extracted from natural language questions, and then aligned with an integer linear programming-based joint inference model to variables in SPARQL queries, used to retrieve answers and additional information from the graphs. Knowledge-based explanations have been presented for visual question-answering (VQA), too. In [82], a combination of DBpedia, WebChild and ConceptNet is used to extract triple patterns that support answers to a given visual question. A combined RNN-LSTM model is used to extract facts from the graphs that are relevant to the input images. Contrary to the current approaches in VQA, this approach allows a form of explicit reasoning, where answers are considered an indicative explanation of the answer. The work extends the one of [83] where reasoning about an image was based on information extracted from a manually-defined subset of DBpedia, and answers were built from pre-defined question templates. An integrated approach is presented by [84], where an embedding space composed of facts from a DBpedia is learnt via a GCN to predict the correct answer given an image and a question about it.

Extending the idea of KB-QA towards conversational AI, [85] uses an automatically-built domain knowledge graph for a conversational-based QA system. The dialog system conducts science dialogues with a user, learning how concepts in a question relate to propositions in facts over a scientific corpus. The learnt concepts and relations are stored in a knowledge graph used to solve questions and explain answers. As the focus is using the graph to facilitate knowledge acquisition, the approach purposefully keeps a simple inferencing mechanism. A shallow inference was also used by [86] in the context of storytelling, building a QA-based conversational agent using knowledge from Framenet, Wordnet and the Open Mind Common-sense dataset. The authors exploit WordNet's verbal causal relationships `cs(Cause, Effect)` and `ent(Action,Consequence)` to generate answers to why questions about a story, as well as generating explanatory sentences. More recently, [87] contrasts the use of patterns, frames, and topic maps as background knowledge in state-of-the-art dialogue systems, employing DBpedia statements as a support for speech recognition. Given an input spoken utterance, the graph is used to identify entities and derive answers from existing triples therein contained. [88] focused on the problem of conversational systems lacking reasoning transparency in the context of smart assistants supporting a user's mundane activities, and showing how this affects critical decision making situations (e.g. privacy, health). The authors suggest that properties of knowledge graphs (e.g. semantic edges and paths between nodes) can be used to provide additional context about a conversation through exploiting a graph's topological features, offering explanations over the system's reasoning. Problems such as scalability when reasoning over knowledge graphs are identified.

The works presented here are summarised in Table 7 and show that knowledge graphs do provide background knowledge for explanations for a variety of contexts such as images, text, speech. The examples in Fig. 6 further reveal a predominant usage of multiple, connected sources rather than single ones. However, one of the main limitations is the limited use of the information in knowledge graphs, mostly restricted to nodes and their closest neighbourhood (i.e. 1-hop relationships such as "a motorbike is smaller than a car"). This is likely due to the computational costs related to a more advanced inferencing mechanisms.

The similarity between *atmosphere* [bn:00006803n] and *ozone* [bn:00060040n] is 2.52 because they are *gas*; they share the same context *chemistry*; they are related to *stratosphere*, *air*, *atmosphere*, *layer*, *ozone*, *atmosphere*, *oxygen*, *gas*.

The similarity between *Boeing* and *plane* is 2.53 because they are related to *airplane*, *aircraft*.

The similarity between *myth* and *satire* is 0.46 because they are *aggregation*, *cosmos*, *cognitive_content*; they are semantically similar to *message*; they form *aggregation*, *division*, *message*, *cosmos*, *cognitive_content*.

(a) Explaining concept similarity through lexical knowledge graphs [76].

Question: What can the red object on the ground be used for ?
Answer: Firefighting
Support Fact: Fire hydrant can be used for fighting fires.

KB	Relationship	#Facts	Examples
DBpedia	Category	35152	(Wii, Category, VideoGameConsole)
ConceptNet	RelatedTo	79789	(Horse, RelatedTo, Zebra), (Wine, RelatedTo, Goblet), (Surfing, RelatedTo, Ocean)
	AtLocation	13683	(Bikini, AtLocation, Beach), (Tap, AtLocation, Bathroom)
	IsA	6011	(Broccoli, IsA, GreenVegetable)
	CapableOf	5837	(Monitor, CapableOf, DisplayImages)
	UsedFor	5363	(Lighthouse, UsedFor, SignalingDanger)
	Desires	3358	(Dog, Desires, PlayFrisbee), (Bee, Desires, Flower)
	HasProperty	2813	(Wedding, HasProperty, Romantic)
	HasA	1665	(Giraffe, HasA, LongTongue), (Cat, HasA, Claw)
	PartOf	762	(RAM, PartOf, Computer), (Tail, PartOf, Zebra)
	ReceivesAction	344	(Books, ReceivesAction, bought at a bookshop)
	CreatedBy	96	(Bread, CreatedBy, Flour), (Cheese, CreatedBy, Milk)
WebChild	Smaller, Better, Slower, Bigger, Taller, ...	38576	(Motorcycle, Smaller, Car), (Apple, Better, VitaminPill), (Train, Slower, Plane), (Watermelon, Bigger, Orange), (Giraffe, Taller, Rhino), (Skating, Faster, Walking)

(b) Supporting facts extracted from ConceptNet and WebChild used to explain answers [82].

Fig. 6. Knowledge-based explanations in natural language applications.

4.5. Predictive and forecasting tasks

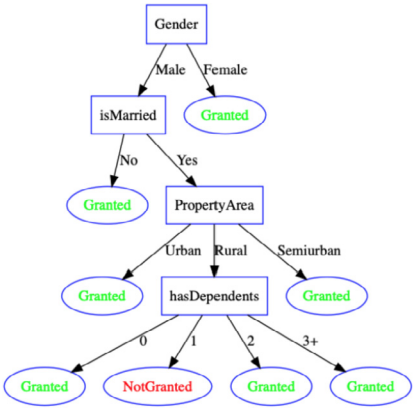
The last body we analyse is the use of knowledge graphs to interpret and explain predictive tasks such as loans applications, market analysis, traffic dynamics etc. These systems rely on the idea that explanations can be derived by linking raw input data points to nodes of the graphs, allowing to retrieve additional information about them through graph navigation (as shown in the examples of Fig. 7). Table 8 summarises the work we include in this area.

In an extended Trepan [89] ontologies-based decision trees are used to explain the outcomes of a prediction model in the context of loan prediction. The user-based evaluation shows that the ontology plays a positive role not only in improving the performance of the model, but also in the perceived understandability of the decisions. A domain ontology is also used by [90], and integrated in a neuro-symbolic architecture with a Restricted Boltzman Machines model in order to explain human behaviours in health social networks. The ontology is used to create symbolic representation for the users and predict their behaviours, while explanations are generated a posteriori step as a set of triples that maximise the likelihood of a user behaviour.

A combination of existing knowledge graphs (Freebase and Wikidata) is instead used by [91] to explain predictions of stock trends using a Temporal Convolutional Network. The graphs are used as external background knowledge to derive embeddings for events and price values extracted from a dataset of financial news. The generated links between events are used as visual explanations to explain unexpected price changes. In their work, [92] uses knowledge graph embeddings

Table 8
Knowledge-based explanations in predictive applications.

	Knowledge graphs					Model				Explanations		
	Type	*Semantics*	*Selection*	*Reusability*	*Number of KG*	*Input*	*Method*	*Task*	*Integration*	*Form*	*Type*	*Interpretability*
[89]	FK	A	man	✗	1	raw	nn	reg	ext	img	cat	pos
[90]	FK	A	man	✗	1	raw	rnn	cls	int	txt	mec	int
[91]	FK	A	man	✓	2	raw	cnn	reg	ext	txt	cat	int
[92]	FK	A	aut	✓	1	raw	kge	cls	int	txt	cat	pos
[93]	CK	T	man	✓	1	txt	cnn	reg	ext	img	cat	pos
[94]	CK	T	man	✓	1	txt	lr/knn	reg	ext	txt	cat	pos



(a) Decision tree using domain ontology concepts to explain the conditions to grant or refuse loans [89].

		Head entity	Relation	Tail entity
Type 1	Predicted triple	Mel Gibson	award nominations	Best Director
	Explanation	Mel Gibson	awards won	Best Director
	Support	Vangelis	award nominations	Best Original Musical
		Vangelis	awards won	Best Original Musical
Type 2	Predicted triple	Aretha Franklin	influenced	Kings of Leon
	Explanation	Kings of Leon	influenced by	Aretha Franklin
	Support	Michael Jackson	influenced	Lady Gaga
		Lady Gaga	influenced by	Michael Jackson
Type 3	Predicted triple	Cayuga County	containedby	New York
	Explanation	Auburn	capital of	Cayuga County
		Auburn	containedby	New York
	Support	Onondaga County	containedby	New York
		Syracuse	capital of	Onondaga County
		Syracuse	containedby	New York

(b) Explanations as supports for predictions made by CrossE [92].

Fig. 7. Knowledge-based explanations for predictive tasks built through graph navigation.

learnt by CrossE to search for explanations of predicted links in a knowledge graph completion tasks. Explanations are regarded as closed paths the head and the tail entities of the predicted link, and the learnt embedding similarity allows to identify the most reliable paths in terms of recall and average support. A knowledge graph-based transfer learning approach is proposed in [93] to explain predictions of delayed flights. The idea is to first learn the predictions based on the dataset and a local OWL ontology, then complementing the learning domain with TBox assertions from DBpedia to explain positive and negative transfers from one domain to another. Similarly, [94] suggests to use RDFS ontologies to enhance input data points of binary classifiers, abstracting them into concepts to be used to derive human-understandable explanations. Concepts are extracted from the DBpedia and the Microsoft Concept Graph and then mapped to a domain ontology.

These approaches integrate knowledge graphs mostly a posteriori, providing explanations about the results of the trained model, and partly ignore the expressiveness of ontology and their reasoning capability to provide explanations for complex models. This aligns with the difficulty of these work in exploiting knowledge graphs to explain models dealing with high-dimensional inputs. A strong focus towards explaining deep nets, while other models (kernel machines, linear or logistic regressions, decision trees) can also become difficult models that could exploit knowledge graphs to provide more transparent results. This was recently brought to the attention by [7], remarking how the field of explainable AI struggles in including knowledge graphs most likely because of the difficulty of deep architectures to integrate prior knowledge.

5. Discussion

Throughout Section 4, we have shown how knowledge graphs have been employed by learning systems of different types and different tasks to provide more meaningful explanations to end users. At first glance, this suggests that the explainable AI community, so far seemingly disconnected from the one of Knowledge Representation and symbolic AI [95], can instead benefit from a number of techniques from the community. Yet, a number of challenges hold and require investigation. Tracing back to our initial research question, the current section provides a summary of the characteristics of existing knowledge-based, explainable (KBX-) systems along with their advantages and limitations.

5.1. What are the characteristics of current knowledge-based explanation systems?

Looking at our analysis, a few interesting observations can be made:

- In domains such as item recommendation or image recognition, the main focus has mostly been the development of KBX-systems providing *mechanistic explanations* for the models' behaviour. This is opposed to data mining contexts, focusing primarily on KBX-systems that could generate *categorical explanations* for the input data. A mix of the two approaches is observed in natural language and forecasting applications, likely due to the variety of tasks to be achieved.
- Similarly, image recognition and recommender systems mostly rely on ABox statements, while the other application domains tend to generate explanations using a combination TBox and ABox statements.
- Areas directly concerned with a user-interaction, e.g. recommender and conversational AI systems, have a strong focus on the integration of knowledge graphs directly in the training model (*model-embedded* knowledge). KBX-systems based on deep architectures are characterised by the integration of symbolic knowledge a posteriori (*post-embedded* knowledge), given the difficulty of integrating large-scale knowledge in deep architectures. More variety is present in image recognition and data mining and knowledge discovery applications.
- A clear distinction on the type of knowledge graphs employed by a KBX-system can be seen depending on the task at hand. Common-sense knowledge graphs are employed to explain behaviours of neural network-based systems for classification tasks such as image recognition and QA, while factual knowledge is employed for prediction and recommendation. The use of domain knowledge graphs is rather restricted to earlier systems to motivate their decision making in rule-based learning.
- Reuse of existing knowledge graphs seems to be an established practice of the latest years, supporting the hypothesis that shared, open knowledge can boost research and reduce development costs. The combination of multiple data sources is somewhat practised, mostly for conversational tasks.
- Extraction of relevant background knowledge from the existing sources is still manually carried across all tasks.

The points above can be summarised as in Fig. 8. The analysed areas are organised across two main axes, respectively indicating the way KBX-systems embed knowledge graphs (model-embedded vs. post-embedded knowledge) and the type of explanation they aim at automatically generating (mechanistic vs. categorical explanations). A third axis representing the type of knowledge graphs used by the systems, is used to colour-code the different areas. A time-based overview of the studies is also shown in Fig. 9. This provides a preliminary systematic overview of the current state of what can be defined the area of knowledge-based, explanation systems.

As a general observation, the use of knowledge graphs to generate behavioural explanations as an a posteriori process still requires investigation. Additionally, a shift from using common-sense knowledge graphs to more domain-specific knowledge graphs can be observed between systems generating model behaviour and data explanations. Similarly, mechanistic explanations tend to be generated using membership assertions (ABox), while categorical explanations are rather generated by exploiting terminological knowledge (TBox). Both trends can be explained by the size of knowledge graphs KBX-systems have to reason upon, i.e. the current KBX-systems rely on complex architectures that are less able to reason over very large knowledge graphs, when compared to systems for categorical explanations, that can reason upon smaller, domain-specific graphs, typically less computationally expensive.

5.2. Knowledge graphs for explainable machine learning: do they work?

In the following, we highlight specific advantages and limitations we have identified in our study. In short:

1. KBX-systems are more understandable as they provide explanations in the form of symbolic, human-readable rules, but need to trade between the structure and succinctness of the generated explanations, which prevents generalisation across tasks;
2. Reusing large-scale knowledge graphs can help the system's accuracy, but techniques for handling noise and efficient knowledge extraction have to be employed;
3. KBX-systems bring reactivity at the costs of computational efficiency.

These points are expanded in the rest of section.

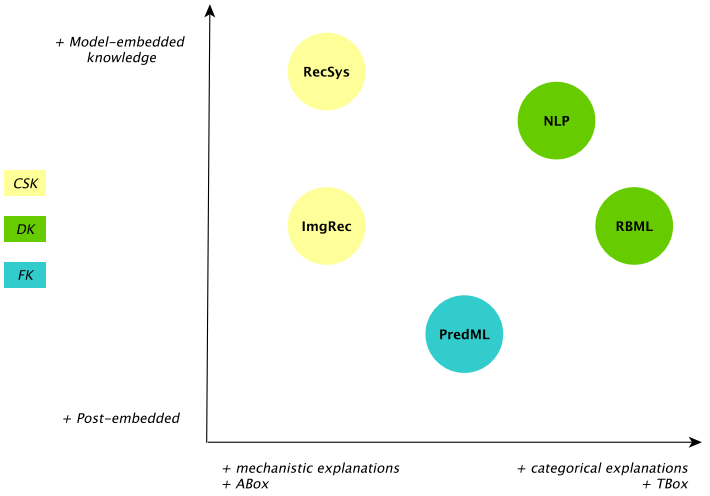


Fig. 8. Overview of existing KBX-systems per application domain (RBML=Rule-based ML, ImgRec=Image Recognition, RecSys=Recommender Systems NLP=Natural Language applications, PredML=Predictive tasks).

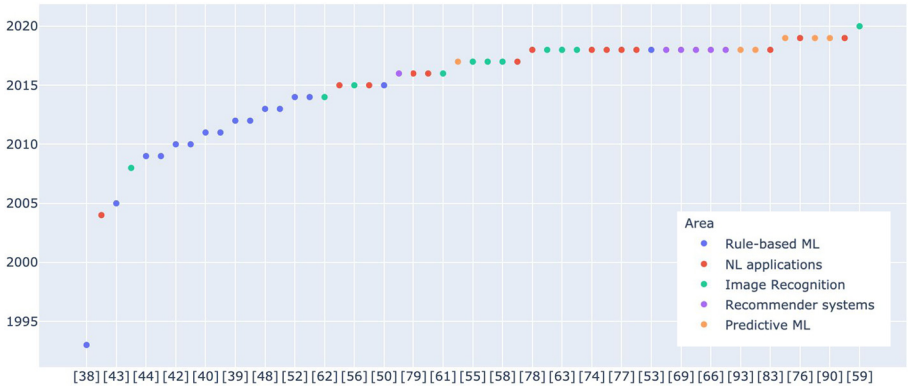


Fig. 9. Time-based overview of the analysed studies, by time and their area. (For interpretation of the colours in the figure(s), the reader is referred to the web version of this article.)

5.2.1. Trading between explanation understandability and task generalisation

KBX-systems express explanations in different forms, spanning from raw texts to natural language explanations to informed visualisations. While there is no agreement of what can be considered a complete and satisfactory explanation, most of the literature consider explanations as answers to a human's needs, and to use knowledge graphs as a mean to augment these answers with additional knowledge a user might not be aware of. Indeed, knowledge graphs provide cross-domain background knowledge that improves the structure of the generated explanations, bringing more informativeness across different tasks; at the same time, the way such knowledge formalisation is closer to the human knowledge conceptualisation, translating into higher trust from humans and thus allowing the application of knowledge graphs to be particularly successful for user-centred applications.

What remains unclear, however, how and to which extent should explanations be structured, i.e. whether they should be composed as strong arguments [36] (cfr. Toulmin's model of argumentations including claims, grounds, warrants and rebuttals), or compact and concise notions. This also highly dependent from the type of background knowledge used in a KBX-system. Systems relying on ABox assertions generate longer explanations for open-domain tasks, but are affected by incompleteness of information and low-reasoning capabilities. Systems relying on TBox allow more structured explanations, but the prohibitive computational costs of reasoning restrict their applicability to domain-/expert-specific contexts. At present, this structure-vs-succinctness trade-off is still highly dependent from the task at hand.

5.2.2. Reusability vs. large-scale management

Reuse of (often multiple) existing knowledge graphs is becoming a common practice likely due to more open, large resources available as well as the ability of system to better scale to them. In this sense, knowledge graphs are an opportunity for to build explainable systems that generate more accurate explanations, also reusable across domains. Managing identity across these sources is here a fundamental requirement. Misalignment between data sources and ambiguity between entities

remains a problem requiring manual filtering and intervention. This is a very well known problem in Knowledge Representation at scale, where data publishing guidelines have been deliberately prioritised to a centralised authority at the costs of an incorrect use of identity in the Semantic Web, sometimes defined as "Identity Crisis" or "the sameAs problem" [96].

Moreover, most of the knowledge-based explanation systems rely on the manual selection of information from the graphs. This choice is mostly motivated by issues related to knowledge graph maintenance, e.g. where information is outdated, missing or incorrect, resulting in the loss of qualitative results and drop of model performance. This represents a substantial limitation of KBX-systems w.r.t. the requirements of explainable AI, one of which consisting in achieving highly performing models that do not require the introduction of any a priori knowledge from the experts. An open issue remains therefore on how to acquire *relevant* knowledge in knowledge graphs, i.e. finding information that is relevant to generate knowledge-based explanations. An open challenge remains therefore to limit the amount of human labour in the knowledge extraction process.

5.2.3. Reactivity vs. approximate reasoning

KBX-systems have exploited knowledge graphs to complement the limitations both of classical and modern (deep) Machine Learning approaches, e.g. need of large training data and inability to transfer the learning task, showing that a flexible solution can support the development of end-to-end approaches generalising across tasks. Additionally, by encoding concepts, relationships and their contexts, knowledge graphs offer an opportunity for systems to integrate inference and causal reasoning, thus improving their reactivity and decision-making ability. This combination is very well-known by neural-symbolic approaches [97], aiming at combining the ability of Machine Learning approaches to learn from experience, and the one of Knowledge Representation frameworks to reason about what has been learnt. In this sense, knowledge graphs allow to develop semantic-interoperable solutions [98], where systems exchange and interpret information produced by different processes in a more efficient way.

Two major limitations emerge here. On the one hand, an evident problem is the one of scalability to very large knowledge graphs, forcing systems to approximate their reasoning and trade between explanations completeness and computational efficiency (at runtime). On the other hand, KBX-systems are still unable to combine information between the learning and reasoning tasks efficiently to achieve complex cognitive tasks (involving perception, decision, and action), also known as the "binding problem" [99]. This results in systems where relating the background information necessary to generate explanations with the training data is a manual task, with data engineers in charge of finding and reusing the relevant information from the graphs, or restricted to the nodes' closest neighbourhood when performed automatically.

6. Current challenges (and ideas to go forward)

Finally, we discuss a set of open challenges that we identified for knowledge-based explainable systems.

Knowledge graph maintenance Explainable AI systems require completeness and accuracy. This means that an important challenge for the field of Knowledge Representation at scale to increase the information coverage and represent more knowledge explicitly across-domains. Additionally, correctness and freshness of the information in large knowledge graph are necessary, requiring not only the investigation of efficient approaches for knowledge graph evolution at scale [100], but also solutions to maintain high-quality cross-domain knowledge graphs without requiring expensive human labour, which can also lead to resources to be discontinued. As already investigated [101], a centralised authoritative hub could be a potential solution to the problem.

Identity management Discrepancy and misalignment between resources of different knowledge graphs is a persistent issue in current KBX-systems. Managing identities is a prerogative for knowledge-based explainable systems to efficiently use the available information and avoid undesirable, wide-ranging effects. While a number of principles exist for publishing and linking resources, a common agreement on what constitutes identical entities is still an open challenge. This also affect the wide-spread adoption of knowledge graphs in eXplainable AI, that cannot tolerated uncertainty over data quality. Solutions to this problems, partly investigated [96], could be services to help data modellers and applications to identify same entities in the real world; better guidelines to correctly use the different types of identity links (e.g. `owl:sameAs`, `owl:equivalentClass`, `skos:exactMatch`). Additionally, error detection and correction approaches to monitor and identify misuse should be investigated [102].

Automated knowledge extraction from graphs Knowledge acquisition from the existing knowledge graphs is still an open challenge which deserved deeper investigation. We believe that there is an urgent need to investigate new heuristics that can deal with the scale of the current knowledge graphs and consequently identify the correct portion information in them automatically. An idea explored in [53,103] but requiring more efforts given the fast-growing nature of knowledge graphs. Application of novel network-analysis methodologies and graph-based strategies to understand the nature of the graphs at hand should also be explored. This would have benefits for KBX-systems in what both the computational costs and the human resource allocation would be significantly reduced.

Understanding the human role An open challenge remains to understand the role of humans in KBX-systems, i.e. if and how much they should be involved in the process of generating explanation. Some KBX-systems have suggested that better performances are achieved when human users provide feedback to the system. In this sense, an idea to be investigated is the applicability of reinforcement learning and humans-in-the-loop in KBX-systems (so far unexplored), integrating principles and methodologies of hybrid intelligence and collaborative AI [104]. Human assessment should also be employed in the development of benchmarks for KBX-systems, currently lacking in the field, to allow a better understanding of what is a "good explanation" from a human perspective. This would also support a better characterisation of useful explanations in terms of types and satisfaction criteria.

From knowledge to meaning Finally, the biggest challenge we mention is the one of capturing meaning. The KBX-systems we analysed exploit knowledge graphs as silos of facts, from which relevant triples are aggregated to support or explain a given observation, without following any particular semantic structure. We argue that knowledge graphs capture much more information beyond simple facts, and that causal, modal or spatio-temporal relationships could be used to building complex narratives such as claims, ideas, stories, behaviours and experiences. By casting simple facts into coherent narratives through semantic models, machines would be able to capture meaning of certain experiences as humans do, and therefore explain the events underlying them more coherently [105]. The study of how information from knowledge graphs can be manipulated and combined to support machines in encompassing meaning will allow the development of a human-centric AI, where machines support human capabilities instead them [106].

7. Conclusions

In this work, we have investigated the use of knowledge graphs in the context of Explainable Machine Learning. Motivated by the idea that a new generation of hybrid, intelligent systems can be developed through combining symbolic (i.e. Knowledge Representation) and subsymbolic (i.e. Machine Learning) methods, we have investigated the hypothesis that machine-readable, large-scale knowledge graphs could be integrated in Explainable Machine Learning systems to provide more meaningful, insightful and trustworthy explanations. We have analysed explainable Machine Learning systems that integrate structured knowledge at large scale in various Machine Learning domains, in order to identify characteristics, strengths and limitations of such an integration. We have provided a comprehensive picture of the current landscape of KBX-systems along with some open research challenges that can be tackled by bringing together the two communities of eXplainable AI and Knowledge Representation, fostering the design of new approaches that efficiently combine very large knowledge graphs and modern learning methods to generate explanations.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] I. Tiddi, M. d'Aquin, E. Motta, An ontology design pattern to define explanations, in: Proceedings of the 8th International Conference on Knowledge Capture, ACM, 2015, p. 3.
[2] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhudinov, R. Zemel, Y. Bengio, Show, attend and tell: neural image caption generation with visual attention, in: International Conference on Machine Learning, 2015, pp. 2048–2057.
[3] M.T. Ribeiro, S. Singh, C. Guestrin, Anchors: high-precision model-agnostic explanations, in: Thirty-Second AAAI Conference on Artificial Intelligence, 2018.
[4] R.M. Byrne, Counterfactuals in explainable artificial intelligence (XAI): evidence from human reasoning, in: Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI-19, 2019, pp. 6276–6282.
[5] F.K. Došilović, M. Brčić, N. Hlupić, Explainable artificial intelligence: a survey, in: 2018 41st International Convention on Information and Communication Technology, Electronics and Microelectronics, MIPRO, IEEE, 2018, pp. 0210–0215.
[6] R. Guidotti, A. Monreale, S. Ruggieri, F. Turini, F. Giannotti, D. Pedreschi, A survey of methods for explaining black box models, ACM Comput. Surv. 51 (5) (2018) 93.
[7] G. Marcus, Deep learning: a critical appraisal, arXiv preprint, arXiv:1801.00631.
[8] F. van Harmelen, A. ten Teije, A boxology of design patterns for hybrid learning and reasoning systems, J. Web Eng. 18 (1) (2019) 97–124.
[9] A. Hogan, E. Blomqvist, M. Cochez, C. d'Amato, G. de Melo, C. Gutierrez, J.E.L. Gayo, S. Kirrane, S. Neumaier, A. Polleres, R. Navigli, A.-C.N. Ngomo, S.M. Rashid, A. Rula, L. Schmelzeisen, J. Sequeda, S. Staab, A. Zimmermann, Knowledge graphs, 2020, arXiv:2003.02320.
[10] O. Hartig, M.T. Özsu, Walking without a map: ranking-based traversal for querying linked data, in: International Semantic Web Conference, Springer, 2016, pp. 305–324.
[11] A. Harth, Link traversal and reasoning in dynamic linked data knowledge bases, Ph.D. thesis, Karlsruher Institut für Technologie (KIT), 2016.
[12] T. Miller, Explanation in artificial intelligence: insights from the social sciences, Artif. Intell. 267 (2019) 1–38, https://doi.org/10.1016/j.artint.2018.07.007, http://www.sciencedirect.com/science/article/pii/S0004370218305988.
[13] O. Biran, C. Cotton, Explanation and justification in machine learning: a survey, in: IJCAI-17 Workshop on Explainable AI (XAI), vol. 8, 2017.
[14] B. Mittelstadt, C. Russell, S. Wachter, Explaining explanations in AI, arXiv preprint, arXiv:1811.01439.
[15] J.S. Mill, A System of Logic, Ratiocinative and Inductive: Being a Connected View of the Principles of Evidence and the Methods of Scientific Investigation, vol. 1, Longmans, Green, and Company, 1884.
[16] C.G. Hempel, P. Oppenheim, Studies in the logic of explanation, Philos. Sci. 15 (2) (1948) 135–175.
[17] B.F. Malle, How the Mind Explains Behavior: Folk Explanations, Meaning, and Social Interaction, MIT Press, 2006.

[18] R. Schank, Explanation Patterns: Understanding Mechanically and Creatively, Psychology Press, 2013.
[19] D. Walton, A dialogue system specification for explanation, Synthese 182 (3) (2011) 349–374.
[20] C. Antaki, I. Leudar, Explaining in conversation: towards an argument model, Eur. J. Soc. Psychol. 22 (2) (1992) 181–194.
[21] H.P. Grice, Logic and conversation, in: Syntax and Semantics, vol. 3, 1975, pp. 41–58.
[22] R.S. Michalski, A Theory and Methodology of Inductive Learning, in: Machine Learning, Springer, 1983, pp. 83–134.
[23] J.A. Overton, Explanation in Science, The University of Western Ontario, 2012.
[24] P.A. Bonatti, S. Decker, A. Polleres, V. Presutti, Knowledge graphs: new directions for knowledge representation on the semantic web (Dagstuhl seminar 18371), Dagstuhl Rep. 8 (9) (2019) 29–111, https://doi.org/10.4230/DagRep.8.9.29, http://drops.dagstuhl.de/opus/volltexte/2019/10328.
[25] G. Malewicz, M.H. Austern, A.J. Bik, J.C. Dehnert, I. Horn, N. Leiser, G. Czajkowski, Pregel: a system for large-scale graph processing, in: Proceedings of the 2010 ACM SIGMOD International Conference on Management of Data, 2010, pp. 135–146.
[26] J.E. Gonzalez, R.S. Xin, A. Dave, D. Crankshaw, M.J. Franklin, I. Stoica, Graphx: graph processing in a distributed dataflow framework, in: 11th {USENIX} Symposium on Operating Systems Design and Implementation ({OSDI} 14), 2014, pp. 599–613.
[27] J. Urbani, C. Jacobs, Adaptive low-level storage of very large knowledge graphs, in: Proceedings of the Web Conference 2020, 2020, pp. 1761–1772.
[28] H. Paulheim, Machine learning with and for semantic web knowledge graphs, in: Reasoning Web International Summer School, Springer, 2018, pp. 110–141.
[29] T. Mitchell, W. Cohen, E. Hruschka, P. Talukdar, B. Yang, J. Betteridge, A. Carlson, B. Dalvi, M. Gardner, B. Kisiel, et al., Never-ending learning, Commun. ACM 61 (5) (2018) 103–115.
[30] N. Noy, Y. Gao, A. Jain, A. Narayanan, A. Patterson, J. Taylor, Industry-scale knowledge graphs: lessons and challenges, Queue 17 (2) (2019) 48–75.
[31] D. Gunning, Explainable artificial intelligence (XAI), Defense Advanced Research Projects Agency, (DARPA), nd Web 2 (2).
[32] A. Páez, The pragmatic turn in explainable artificial intelligence (xai), Minds Mach. 29 (3) (2019) 441–459.
[33] R. Caruana, Y. Lou, J. Gehrke, P. Koch, M. Sturm, N. Elhadad, Intelligible models for healthcare: predicting pneumonia risk and hospital 30-day readmission, in: Proceedings of the 21st ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2015, pp. 1721–1730.
[34] S. Chari, O. Seneviratne, D.M. Gruen, M.A. Foreman, A.K. Das, D.L. McGuinness, Explanation ontology: a model of explanations for user-centered AI, in: International Semantic Web Conference, Springer, 2020, pp. 228–243.
[35] M.Y. Jaradeh, A. Oelen, K.E. Farfar, M. Prinz, J. D'Souza, G. Kismihók, M. Stocker, S. Auer, Open research knowledge graph: next generation infrastructure for semantic scholarly knowledge, in: Proceedings of the 10th International Conference on Knowledge Capture, 2019, pp. 243–246.
[36] S. Gregor, I. Benbasat, Explanations from intelligent systems: theoretical foundations and implications for practice, MIS Q. (1999) 497–530.
[37] R.W. Southwick, Explaining reasoning: an overview of explanation in knowledge-based systems, Knowl. Eng. Rev. 6 (1) (1991) 1–19.
[38] U. Fayyad, G. Piatetsky-Shapiro, P. Smyth, From data mining to knowledge discovery in databases, AI Mag. 17 (3) (1996) 37.
[39] G.G. Towell, J.W. Shavlik, Extracting refined rules from knowledge-based neural networks, Mach. Learn. 13 (1) (1993) 71–101.
[40] T.A. Russ, C. Ramakrishnan, E.H. Hovy, M. Bota, G.A. Burns, Knowledge engineering tools for reasoning with scientific observations and interpretations: a neural connectivity use case, BMC Bioinform. 12 (1) (2011) 351.
[41] M. d'Aquin, G. Kronberger, M.C. Suárez-Figueroa, Combining data mining and ontology engineering to enrich ontologies and linked data, KNOW@ LOD 868 (2012) 19–24.
[42] M.A. Domingues, S.O. Rezende, Using taxonomies to facilitate the analysis of the association rules, arXiv preprint, arXiv:1112.1734.
[43] V. Svátek, J. Rauch, M. Ralbovskỳ, Ontology-enhanced association mining, in: Semantics, Web and Mining, Springer, 2005, pp. 163–179.
[44] C. Marinica, F. Guillet, Knowledge-based interactive postmining of association rules using ontologies, IEEE Trans. Knowl. Data Eng. 22 (6) (2010) 784–797.
[45] P.K. Novak, A. Vavpetic, I. Trajkovski, N. Lavrac, Towards semantic data mining with g-segs, in: Proceedings of the 11th International Multiconference, Information Society, IS, 2009.
[46] A. Vavpetič, V. Podpečan, N. Lavrač, Semantic subgroup explanations, J. Intell. Inf. Syst. 42 (2) (2014) 233–254.
[47] Z. Huang, H. Chen, T. Yu, H. Sheng, Z. Luo, Y. Mao, Semantic text mining with linked data, in: 2009 Fifth International Joint Conference on INC, IMS and IDC, IEEE, 2009, pp. 338–343.
[48] M. d'Aquin, N. Jay, Interpreting data mining results with linked data for learning analytics: motivation, case study and directions, in: Proceedings of the Third International Conference on Learning Analytics and Knowledge, ACM, 2013, pp. 155–164.
[49] N. Jay, M. d'Aquin, Linked data and online classifications to organise mined patterns in patient data, in: AMIA Annual Symposium Proceedings, vol. 2013, American Medical Informatics Association, 2013, p. 681.
[50] H. Paulheim, Generating possible interpretations for statistics from linked open data, in: Extended Semantic Web Conference, Springer, 2012, pp. 560–574.
[51] P. Ristoski, H. Paulheim, Visual analysis of statistical data on maps using linked open data, in: International Semantic Web Conference, Springer, 2015, pp. 138–143.
[52] V. Mulwad, T. Finin, Z. Syed, A. Joshi, et al., Using linked data to interpret tables, in: Proceedings of the First International Workshop on Consuming Linked Data, 2010.
[53] I. Tiddi, M. d'Aquin, E. Motta, Dedalo: looking for clusters explanations in a labyrinth of linked data, in: European Semantic Web Conference, Springer, 2014, pp. 333–348.
[54] A. Campero, A. Pareja, T. Klinger, J. Tenenbaum, S. Riedel, Logical rule induction and theory learning using neural theorem proving, arXiv preprint, arXiv:1809.02193.
[55] N.E. Maillot, M. Thonnat, Ontology based complex object recognition, Image Vis. Comput. 26 (1) (2008) 102–113.
[56] R.T. Icarte, J.A. Baier, C. Ruz, A. Soto, How a general-purpose commonsense ontology can improve performance of learning-based image retrieval, arXiv preprint, arXiv:1705.08844.
[57] V. Ordonez, W. Liu, J. Deng, Y. Choi, A.C. Berg, T.L. Berg, Predicting entry-level categories, Int. J. Comput. Vis. 115 (1) (2015) 29–43.
[58] Q. Liao, T. Poggio, Object-oriented deep learning, Tech. rep., Center for Brains, Minds and Machines (CBMM), 2017.
[59] M.K. Sarker, N. Xie, D. Doran, M. Raymer, P. Hitzler, Explaining trained neural networks with semantic web technologies: first steps, arXiv preprint, arXiv:1710.04324.
[60] Z.A. Daniels, L.D. Frank, C.J. Menart, M. Raymer, P. Hitzler, A framework for explainable deep neural models using external knowledge graphs, in: Artificial Intelligence and Machine Learning for Multi-Domain Operations Applications II, vol. 11413, International Society for Optics Photonics, 2020, p. 114131C.
[61] M. Alirezaie, M. Längkvist, M. Sioutis, A. Loutfi, A symbolic approach for explaining errors in image classification tasks, in: International Joint Conference on Artificial Intelligence (IJCAI), 2018.
[62] K. Marino, R. Salakhutdinov, A. Gupta, The more you know: using knowledge graphs for image classification, arXiv preprint, arXiv:1612.04844.
[63] Y. Zhu, A. Fathi, L. Fei-Fei, Reasoning about object affordances in a knowledge base representation, in: European Conference on Computer Vision, Springer, 2014, pp. 408–424.
[64] X. Wang, Y. Ye, A. Gupta, Zero-shot recognition via semantic embeddings and knowledge graphs, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 6857–6866.

[65] F.Å. Nielsen, Linking imagenet wordnet synsets with wikidata, arXiv preprint, arXiv:1803.04349.
[66] H. Wang, F. Zhang, X. Xie, M. Guo, Dkn: deep knowledge-aware network for news recommendation, in: Proceedings of the 2018 World Wide Web Conference on World Wide Web, International World Wide Web Conferences Steering Committee, 2018, pp. 1835–1844.
[67] Q. Ai, V. Azizi, X. Chen, Y. Zhang, Learning heterogeneous knowledge base embeddings for explainable recommendation, Algorithms 11 (9) (2018) 137.
[68] V. Bellini, A. Schiavone, T. Di Noia, A. Ragone, E. Di Sciascio, Knowledge-aware autoencoders for explainable recommender systems, in: Proceedings of the 3rd Workshop on Deep Learning for Recommender Systems, ACM, 2018, pp. 24–31.
[69] V. Bellini, T. Di Noia, E. Di Sciascio, A. Schiavone, Semantics-aware autoencoder, IEEE Access 7 (2019) 166122–166137.
[70] J. Huang, W.X. Zhao, H. Dou, J.-R. Wen, E.Y. Chang, Improving sequential recommendation with knowledge-enhanced memory networks, in: The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval, ACM, 2018, pp. 505–514.
[71] C. Musto, F. Narducci, P. Lops, M. De Gemmis, G. Semeraro, Explod: a framework for explaining recommendations based on the linked open data cloud, in: Proceedings of the 10th ACM Conference on Recommender Systems, ACM, 2016, pp. 151–154.
[72] V. Lully, P. Laublet, M. Stankovic, F. Radulovic, Enhancing explanations in recommender systems with knowledge graphs, Proc. Comput. Sci. 137 (2018) 211–222.
[73] H. Paulheim, A. Gangemi, Serving dbpedia with dolce – more than just adding a cherry on top, in: M. Arenas, O. Corcho, E. Simperl, M. Strohmaier, M. d'Aquin, K. Srinivas, P. Groth, M. Dumontier, J. Heflin, K. Thirunarayan, S. Staab (Eds.), The Semantic Web - ISWC 2015, Springer International Publishing, Cham, 2015, pp. 180–196.
[74] D.J. Hilton, Conversational processes and causal explanation, Psychol. Bull. 107 (1) (1990) 65.
[75] T. Mihaylov, A. Frank, Knowledgeable reader: enhancing cloze-style reading comprehension with external commonsense knowledge, arXiv preprint, arXiv:1805.07858.
[76] D. Colla, E. Mensa, D.P. Radicioni, A. Lieto, Tell me why: computational explanation of conceptual similarity judgments, in: International Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems, Springer, 2018, pp. 74–85.
[77] V. Silva, A. Freitas, S. Handschuh, Exploring knowledge graphs in an interpretable composite approach for text entailment, in: Thirty-Third AAAI Conference on Artificial Intelligence, AAAI Press, 2019.
[78] R. Musa, X. Wang, A. Fokoue, N. Mattei, M. Chang, P. Kapanipathi, B. Makni, K. Talamadupula, M. Witbrock, Answering science exam questions using query reformulation with background knowledge, in: Automated Knowledge Base Construction (AKBC), 2018.
[79] W. Zhong, D. Tang, N. Duan, M. Zhou, J. Wang, J. Yin, Improving question answering by commonsense-based pre-training, arXiv preprint, arXiv: 1809.03568.
[80] Z. Dai, L. Li, W. Xu, Cfo: conditional focused neural question answering with large-scale knowledge bases, arXiv preprint, arXiv:1606.01994.
[81] Y. Zhang, S. He, K. Liu, J. Zhao, A joint model for question answering over multiple knowledge bases, in: Thirtieth AAAI Conference on Artificial Intelligence, 2016.
[82] P. Wang, Q. Wu, C. Shen, A. Dick, A. van den Hengel, Fvqa: fact-based visual question answering, IEEE Trans. Pattern Anal. Mach. Intell. 40 (10) (2018) 2413–2427.
[83] P. Wang, Q. Wu, C. Shen, A. v. d. Hengel, A. Dick, Explicit knowledge-based reasoning for visual question answering, arXiv preprint, arXiv:1511.02570.
[84] M. Narasimhan, S. Lazebnik, A. Schwing, Out of the box: reasoning with graph convolution nets for factual visual question answering, in: Advances in Neural Information Processing Systems, 2018, pp. 2654–2665.
[85] B. Hixon, P. Clark, H. Hajishirzi, Learning knowledge graphs for question answering through conversational dialog, in: Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2015, pp. 851–861.
[86] P. Tarau, E. Figa, Knowledge-based conversational agents and virtual storytelling, in: Proceedings of the 2004 ACM Symposium on Applied Computing, ACM, 2004, pp. 39–44.
[87] A.J. Kumar, S. Auer, C. Schmidt, et al., Towards a knowledge graph based speech interface, arXiv preprint, arXiv:1705.09222.
[88] S. Hepenstal, N. Kodagoda, L. Zhang, P. Paudyal, B.L. Wong, Algorithmic transparency of conversational agents, in: CEUR Workshop Proceedings, 2019.
[89] R. Confalonieri, F.M. del Prado, S. Agramunt, D. Malagarriga, D. Faggion, T. Weyde, T.R. Besold, An ontology-based approach to explaining artificial neural networks, arXiv preprint, arXiv:1906.08362.
[90] N. Phan, D. Dou, H. Wang, D. Kil, B. Piniewski, Ontology-based deep learning for human behavior prediction with explanations in health social networks, Inf. Sci. 384 (2017) 298–313.
[91] S. Deng, N. Zhang, W. Zhang, J. Chen, J.Z. Pan, H. Chen, Knowledge-driven stock trend prediction and explanation via temporal convolutional network, in: Companion Proceedings of the 2019 World Wide Web Conference, 2019, pp. 678–685.
[92] W. Zhang, B. Paudel, W. Zhang, A. Bernstein, H. Chen, Interaction embeddings for prediction and explanation in knowledge graphs, in: Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining, 2019, pp. 96–104.
[93] J. Chen, F. Lecue, J.Z. Pan, I. Horrocks, H. Chen, Knowledge-based transfer learning explanation, in: Sixteenth International Conference on Principles of Knowledge Representation and Reasoning, 2018.
[94] F. Lécué, J. Wu, Semantic explanations of predictions, arXiv preprint, arXiv:1805.10587.
[95] J.M. Alonso, C. Castiello, C. Mencar, A bibliometric analysis of the explainable artificial intelligence research field, in: International Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems, Springer, 2018, pp. 3–15.
[96] J. Raad, W. Beek, F. Van Harmelen, N. Pernelle, F. Saïs, Detecting erroneous identity links on the web using network metrics, in: International Semantic Web Conference, Springer, 2018, pp. 391–407.
[97] S. Bader, P. Hitzler, Dimensions of neural-symbolic integration-a structured survey, arXiv preprint cs/0511042.
[98] M. d'Aquin, N.F. Noy, Where to publish and find ontologies? A survey of ontology libraries, J. Web Semant. 11 (2012) 96–111.
[99] J. Feldman, The neural binding problem(s), Cogn. Neurodyn. 7 (1) (2013) 1–11.
[100] H. Paulheim, Knowledge graph refinement: a survey of approaches and evaluation methods, Semant. Web 8 (3) (2017) 489–508.
[101] W. Beek, L. Rietveld, S. Schlobach, F. van Harmelen, Lod laundromat: why the semantic web needs centralization (even if we don't like it), IEEE Internet Comput. 20 (2) (2016) 78–81.
[102] I. Tiddi, M. d'Aquin, E. Motta, Quantifying the bias in data links, in: International Conference on Knowledge Engineering and Knowledge Management, Springer, 2014, pp. 531–546.
[103] I. Tiddi, M. d'Aquin, E. Motta, Learning to assess linked data relationships using genetic programming, in: International Semantic Web Conference, Springer, 2016, pp. 581–597.
[104] Y. Gil, J. Honaker, S. Gupta, Y. Ma, V. D'Orazio, D. Garijo, S. Gadewar, Q. Yang, N. Jahanshad, Towards human-guided machine learning, in: Proceedings of the 24th International Conference on Intelligent User Interfaces, ACM, 2019, pp. 614–624.
[105] L. Steels, Personal dynamic memories are necessary to deal with meaning and understanding in human-centric AI, in: NeHuAI@ ECAI, 2020, pp. 11–16.
[106] Z. Akata, D. Balliet, M. De Rijke, F. Dignum, V. Dignum, G. Eiben, A. Fokkens, D. Grossi, K. Hindriks, H. Hoos, et al., A research agenda for hybrid intelligence: augmenting human intellect with collaborative, adaptive, responsible, and explainable artificial intelligence, Computer 53 (08) (2020) 18–28.