A Gentle Introduction to Knowledge Graphs

Navdeep Kaur

Research Associate

nkaur@turing.ac.uk

Me @Turing

Project:

Robust Inference with Probabilistic Answer Set Programs Scaffolds for Large Language Models

Organisers

Dr Pranava Madhyastha

Professor Alessandra Russo

Professor Anthony Cohn
Foundational Models Theme lead

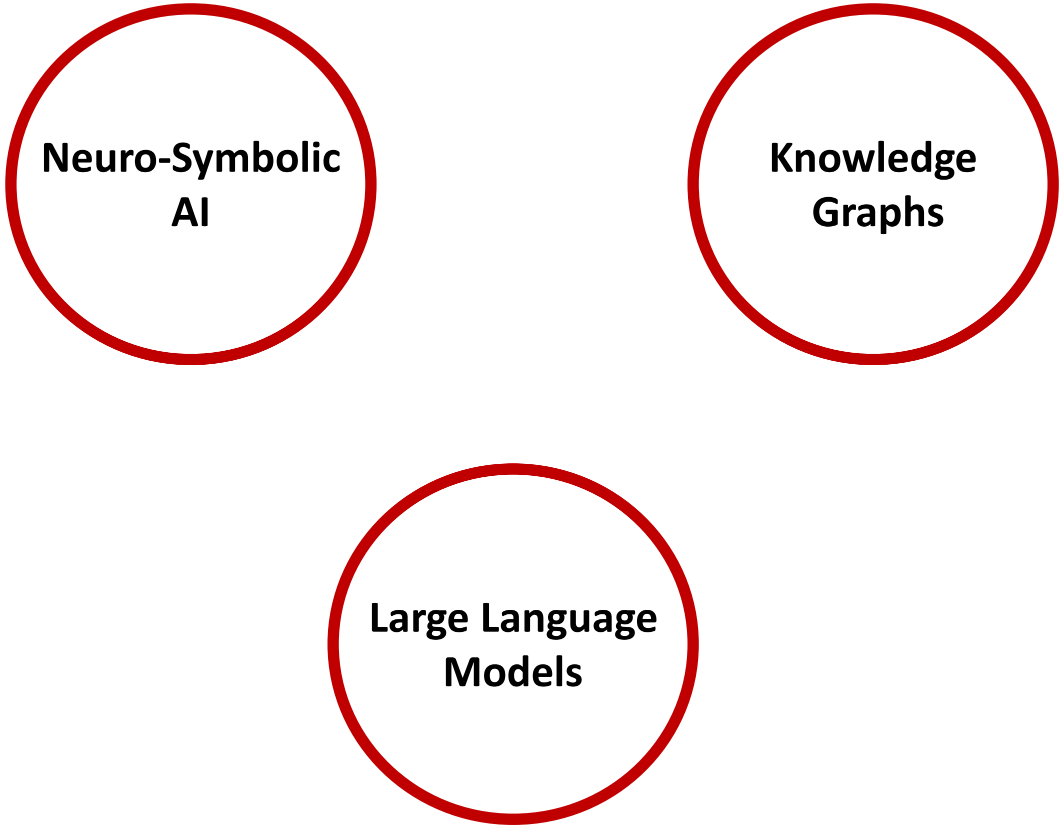
Ruth Drysdale
Programme Manager, Artificial Intelligence (AI)

Chiara Sbruazzo
Research Project Manager, Fundamental AI

Dr. Navdeep Kaur
Research Associate

Dr. Lachlan McPheat

My Research Interests



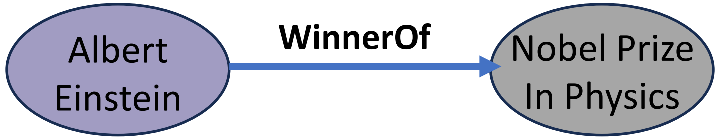
Outline

- **Knowledge Graphs**
- Knowledge Graph Completion
- Knowledge Graph Completion – the “spin offs”
- Knowledge Graphs and LLMs

Knowledge Graphs

- A **Knowledge Graph** is a structured representation of knowledge that captures the relationships between different entities in a way that can be easily understood and processed by both humans and machines.
- Knowledge Graphs are used to model real-world information and concepts by organizing them as a network of entities and the relationships between them.
- Formally, **Knowledge Graph (KG)** is a directed, labeled graph $\mathbf{G} = (\mathbf{V}, \mathbf{E}, \mathbf{R})$ where:
- $\mathbf{V}$ is a set of **vertices** (or nodes), representing entities or concepts. Each vertex $v \in \mathbf{V}$ typically corresponds to an entity:
  - such as a person, place, object, event, or abstract concept
- $\mathbf{E}$ is a set of **edges** (or links), representing the relationships between entities. Each edge $e = (v_i, r, v_j) \in \mathbf{E}$ is a directed relationship from entity v_i to entity v_j with r as the relationship type. $e = (v_i, r, v_j) \in \mathbf{E}$ Is usually represented as:
  - (h, r, t), (head, relation, tail)
  - (s, r, o) (subject, relation, object)
- $\mathbf{R}$ is a set of **relationship types** (also known as predicates or properties), which define the nature of the relationships between entities.

(Albert Einstein, **WinnerOf**, Nobel Prize in Physics)



https://www.youtube.com/watch?v=mmQl6VGvX-c

Freebase

(Albert Einstein, **SonOf**, Hermann Einstein)

Albert Einstein, **BornIn**, German Empire)

(Albert Einstein, **GraduatedFrom**, University of Zurich)

(Albert Einstein, **WinnerOf**, Nobel Prize in Physics)

(Albert Einstein, **ExpertIn**, Physics)

(Nobel Prize in Physics, **AwardIn**, Physics)

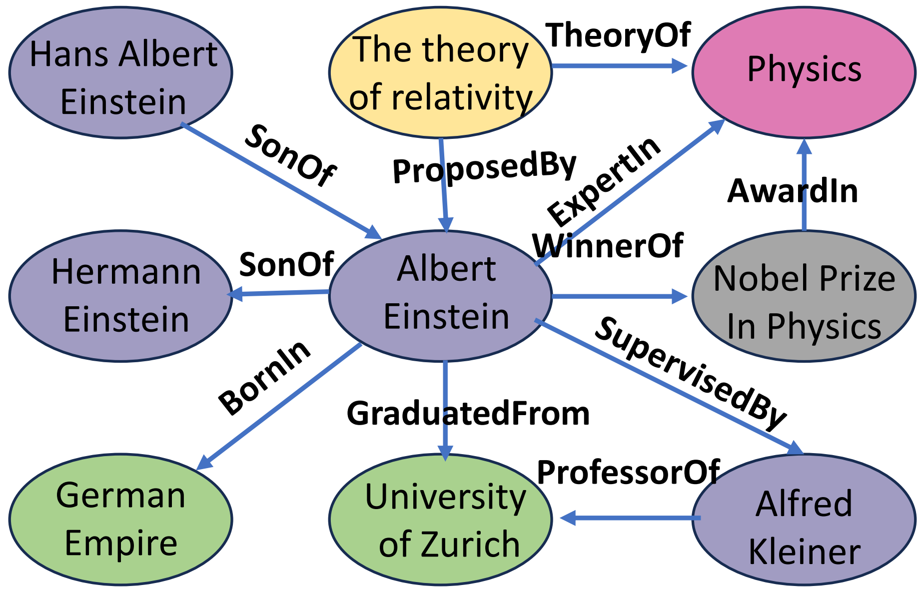
(The Theory of Relativity, **TheoryOf**, Physics)

(Albert Einstein, **SupervisedBy**, Alfred Kleiner)

(Alfred Kleiner, **ProfessorOf**, University of Zurich)

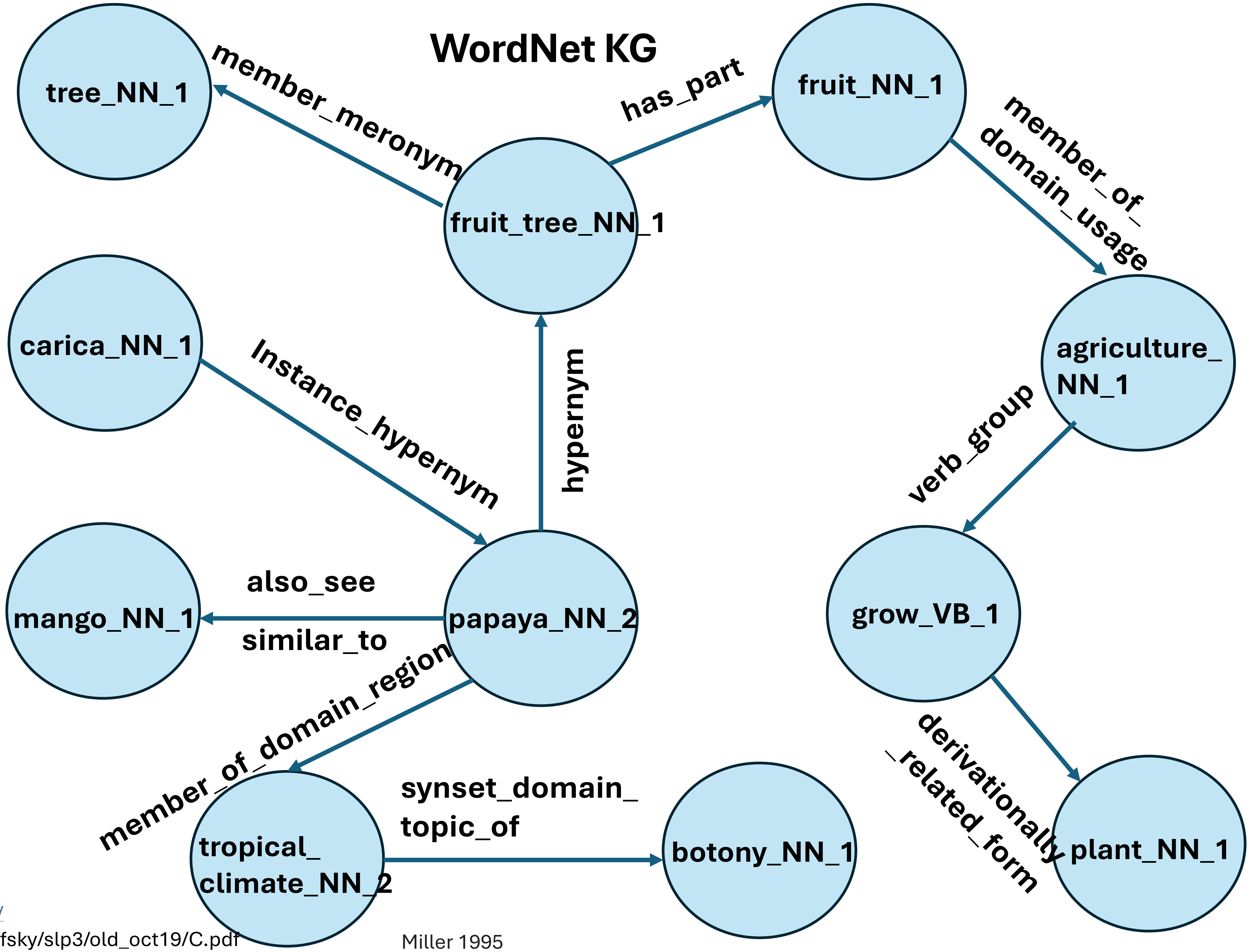
(The theory of relativity, **ProposedBy**, Albert Einstein)

(Hans Albert Einstein, **SonOf**, Albert Einstein)



Bollacker et al. 2008

WordNet KG



https://wordnet.princeton.edu/
https://web.stanford.edu/~jurafsky/slp3/old_oct19/C.pdf

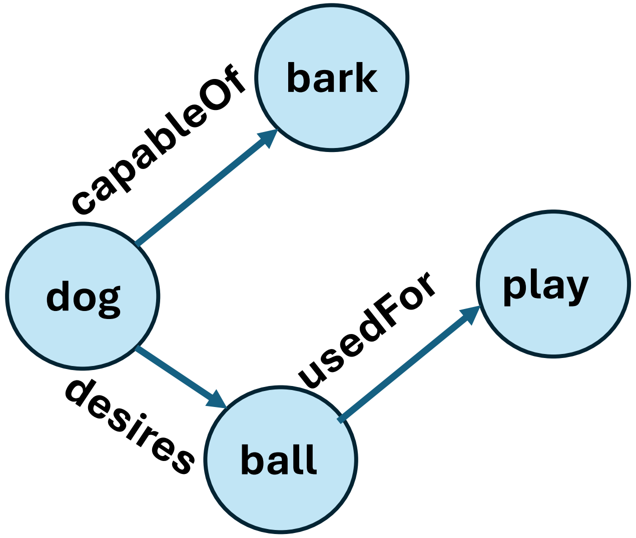
Miller 1995

Common Sense Knowledge Graphs

Common Sense Knowledge Graph

- ConceptNet is a KG that connects the words and the phrases in natural language with labelled edges
- **Nodes**: represent everyday concepts like "Dog," "Bark," "Leash," "Owner," etc.
- **Edges**: represent the common-sense relationships between these concepts
- The goal is to improve the natural language applications by allowing applications to better understand the meaning behind the words people use.

- ConceptNet 5.5[1]
    - 21M edges, 8M nodes
    - 36 Relations
    - Multilingual KG (83 languages)
- Examples:
    - (a net, usedFor, fishing): a net is used for catching fishing
    - ("leaves", formOf, "leaf"): "leaves" is a form of the word "word"
    - (dog, hasA, tail): dog has a tail
    - (wheel, partOf, car): a wheel is a part of a car



1. https://github.com/commonsense/conceptnet5/wiki/Relations
2. Concept 5.5: An Open Multilingual Graph of General Knowledge, AAAI 2017

Knowledge Graphs

TABLE VII: Statistics of datasets with general knowledge when originally released

Dataset	# entities	# facts	Website
WordNet [234]	117,597	207,016	https://wordnet.princeton.edu
OpenCyc [235]	47,000	306,000	https://www.cyc.com/opencyc/
Cyc [235]	~250,000	~2,200,000	https://www.cyc.com
YAGO [237]	1,056,638	~5,000,000	http://www.mpii.mpg.de/~suchanek/yago
DBpedia [236]	~1,950,000	~103,000,000	https://wiki.dbpedia.org/develop/datasets
Freebase [238]	-	~125,000,000	https://developers.google.com/freebase/
NELL [73]	-	242,453	http://rtw.ml.cmu.edu/rtw/
Wikidata [239]	14,449,300	30,263,656	https://www.wikidata.org/wiki
Probase IsA	12,501,527	85,101,174	https://concept.research.microsoft.com/Home/Download
Google KG	$>$ 500 million	$>$ 3.5 billion	https://developers.google.com/knowledge-graph

A Survey on Knowledge Graphs: Representation, Acquisition and Application, Ji et al. 2021, IEEE TRANSACTIONS ON NEURAL NETWORKS AND LEARNING SYSTEMS

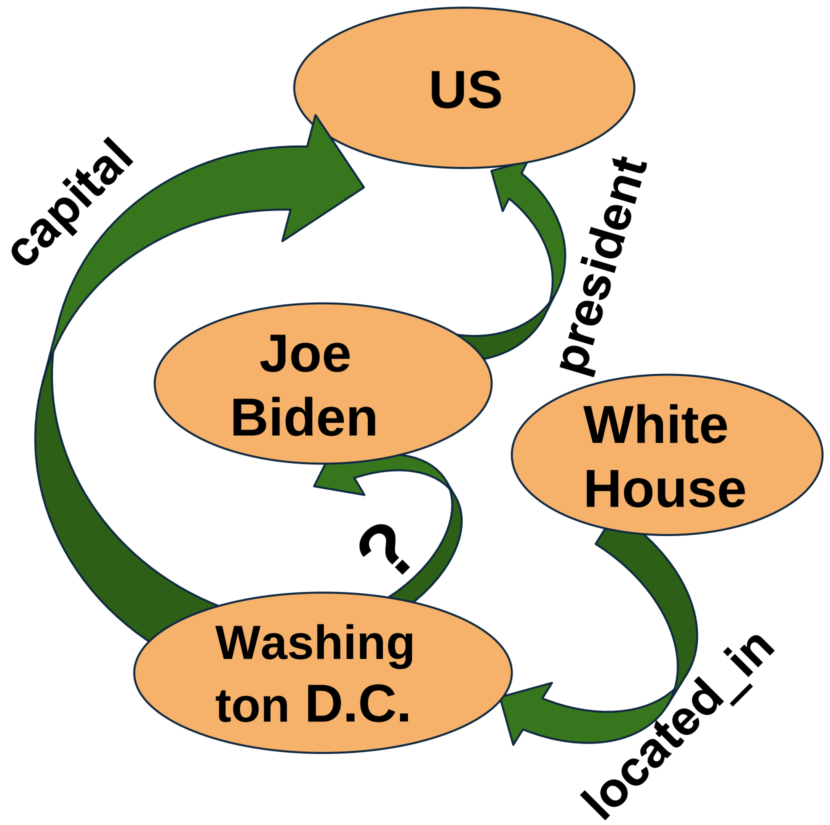
Miller 1995 (WORDNET); Suchanek et al. 2007 (YAGO); Bollacker et al. 2008 (FREEBASE); Carlson et al. 2010 (NELL); Lehmann et al. 2014 (DBPEDIA); Dong et al. 2014 (GOOGLE).

Outline

- Knowledge Graphs
- **Knowledge Graph Completion**
- Knowledge Graph Completion – the “spin offs”
- Knowledge Graphs and LLMs

Knowledge Graph Completion

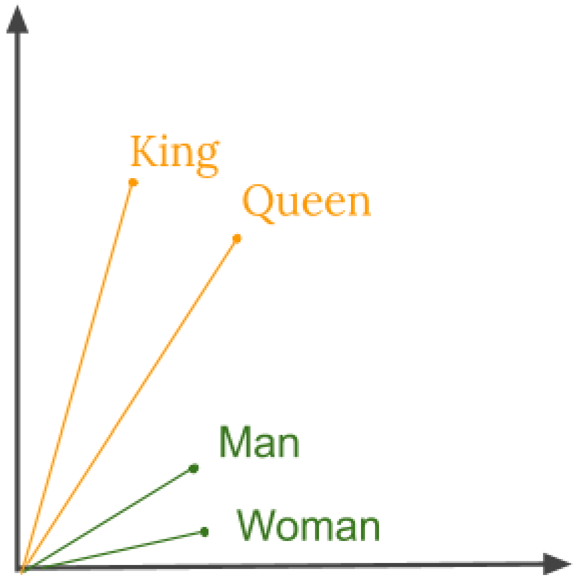
- These knowledge graphs are ever-expanding with newer facts being added to them everyday
- Most of these knowledge graphs are incomplete and have important links missing in them.
- The problem of enriching KGs by inferring missing links between existing entities of KG is formulated as a **KG Completion (KGC)** task
- KG Completion (KGC) ≡ Link Prediction Problem



Knowledge Graph

Miller 1995; Suchanek et al. 2007; Bollacker et al. 2008; Carlson et al. 2010; Lehmann et al. 2014; Dong et al. 2014.

Word2Vec: the vector representation of words



- In word embeddings (word2vec), all individual words are represented as real-valued vectors in a predefined, n-dimensional, vector space.
- Each word is mapped to one vector and the vector values are learned in a way that resembles a neural network
- Two semantically similar words are represented by almost similar vectors that are very closely placed in a vector space.

Efficient Estimation of Word Representations in Vector Space, Tomas Mikolov, Kai Chen, Greg Corrado, Jeffrey Dean, 2013
Figure from: https://www.mygreatlearning.com/blog/word-embedding/

Knowledge Graphs Embeddings

- head, relation, tail are represented by learnable vectors in R^d
- TransE Model:
  - Intuition:
    - Embedding of relation r corresponds to the translation between h and t i.e. **h + r ≈ t** for a given triple (h,r,t)
  - The dissimilarity measure of TransE is defined as:

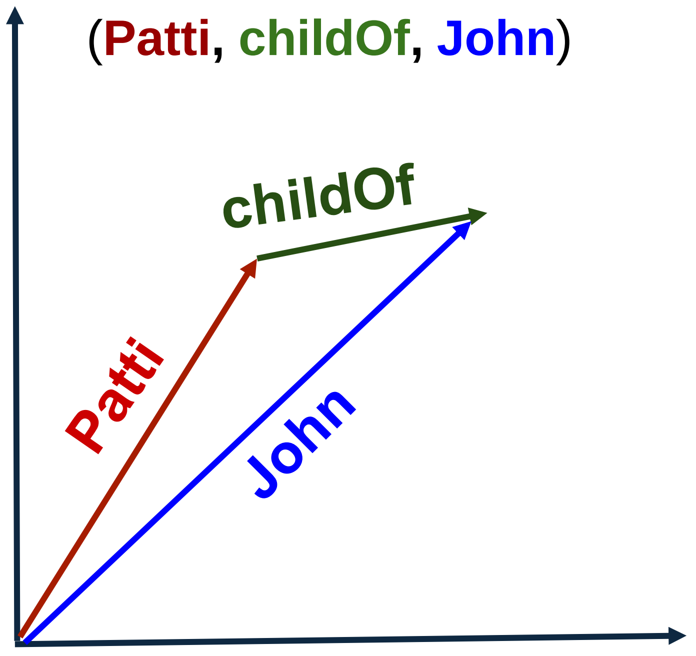
$$d(h,r,t) = ||\mathbf{h} + \mathbf{r} - \mathbf{t}||_2^2$$

  - The embeddings or vectors are optimized by margin-based ranking function:

$$\sum_{pos} \sum_{neg \in S'} \left[\gamma + ||\mathbf{h} + \mathbf{r} - \mathbf{t}||_2^2 - ||\mathbf{h'} + \mathbf{r'} - \mathbf{t'}||_2^2\right]_+$$

where $[x]_+ = \max\{0, x\}$

$$S' = \{(h', r, t) \mid h' \in E\} \cup \{(h, r, t') \mid t' \in E\} \cup \{(h, r', t) \mid r' \in R\}$$



Patti	childOf	John
0.036	0.102	0.138
-0.120	0.671	0.551
…	…	…
0.323	-0.101	0.222

Translating embeddings for modeling multi-relational data, Bordes et al. 2013

Knowledge Graph Completion: A decade's (2011-2020) worth of work

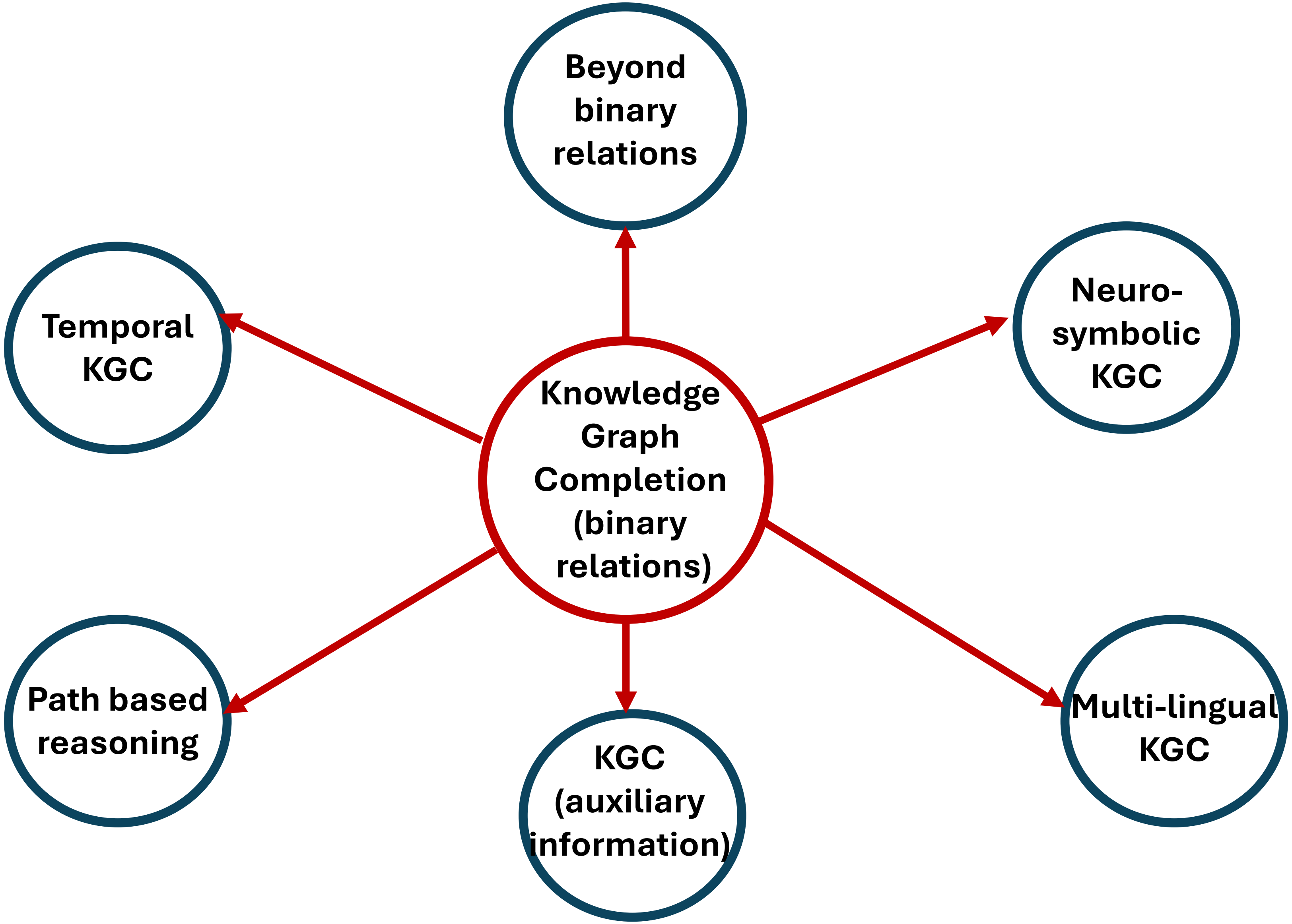
Knowledge Graph Completion Models

A Survey on Knowledge Graphs: Representation, Acquisition and Applications, IEEE TRANSACTIONS ON NEURAL NETWORKS AND LEARNING SYSTEMS, 2021

Outline

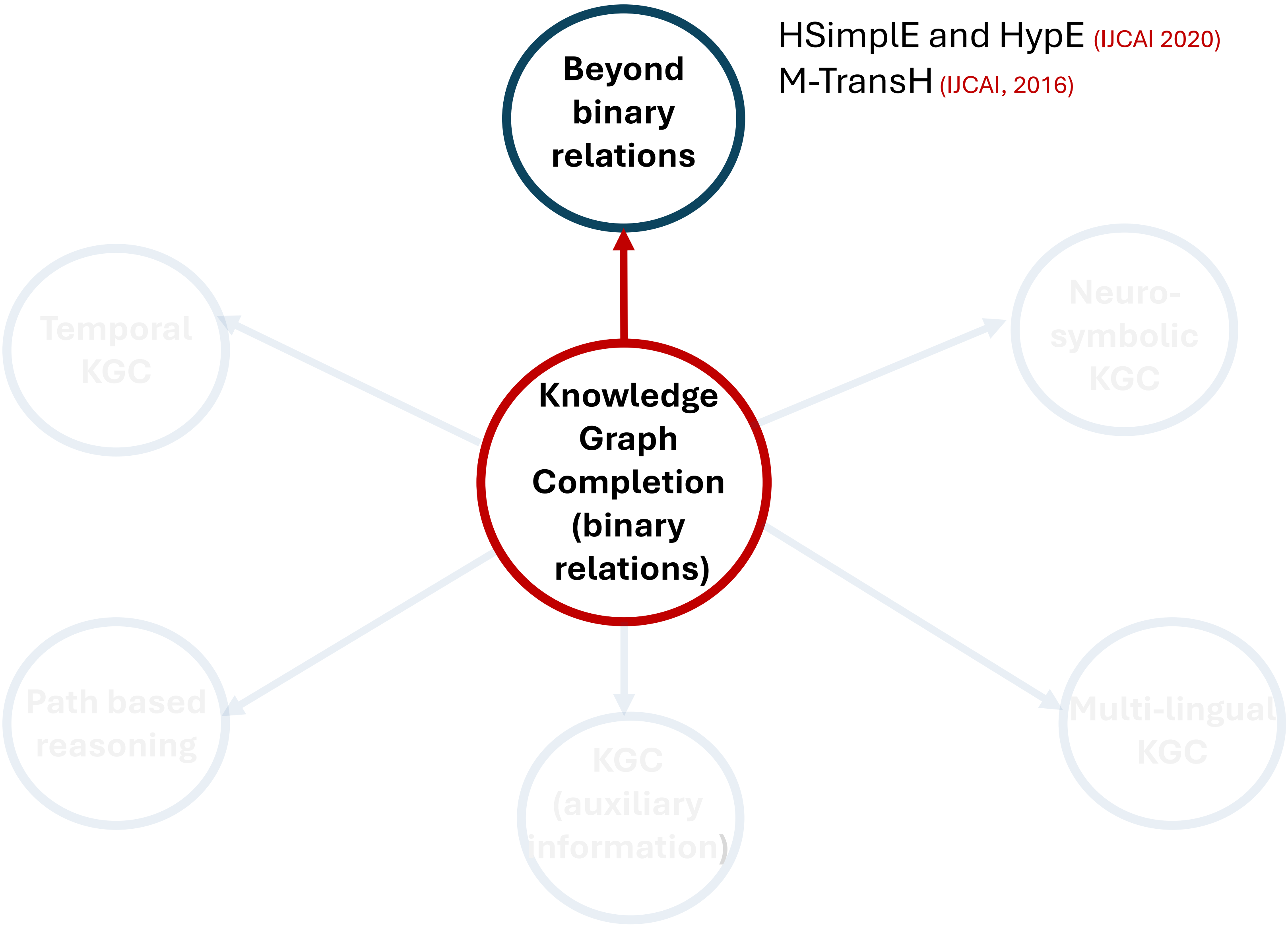
- Knowledge Graphs
- Knowledge Graph Completion
- **Knowledge Graph Completion – the “spin offs”**
- Knowledge Graphs and LLMs

Knowledge Graph Completion – the “spin-offs”



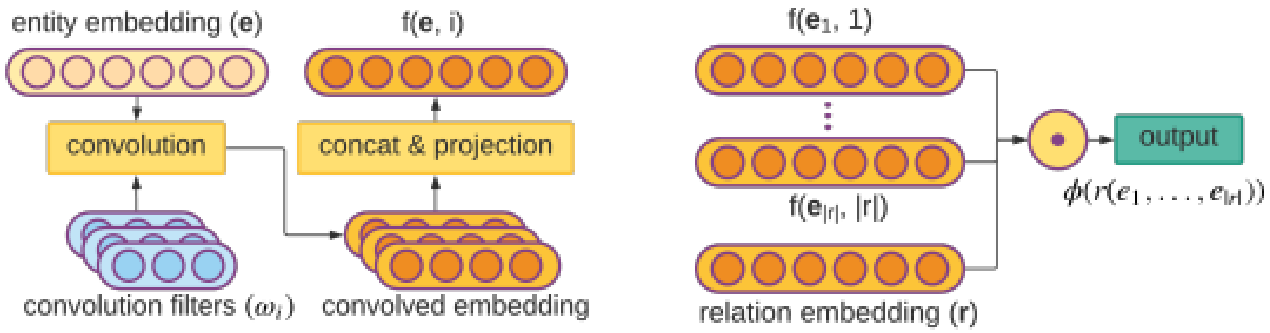
Knowledge Graph Completion – the "spin-offs"

HSimplE and HypE (IJCAI 2020)
M-TransH (IJCAI, 2016)



Knowledge Hyper-Graph Completion

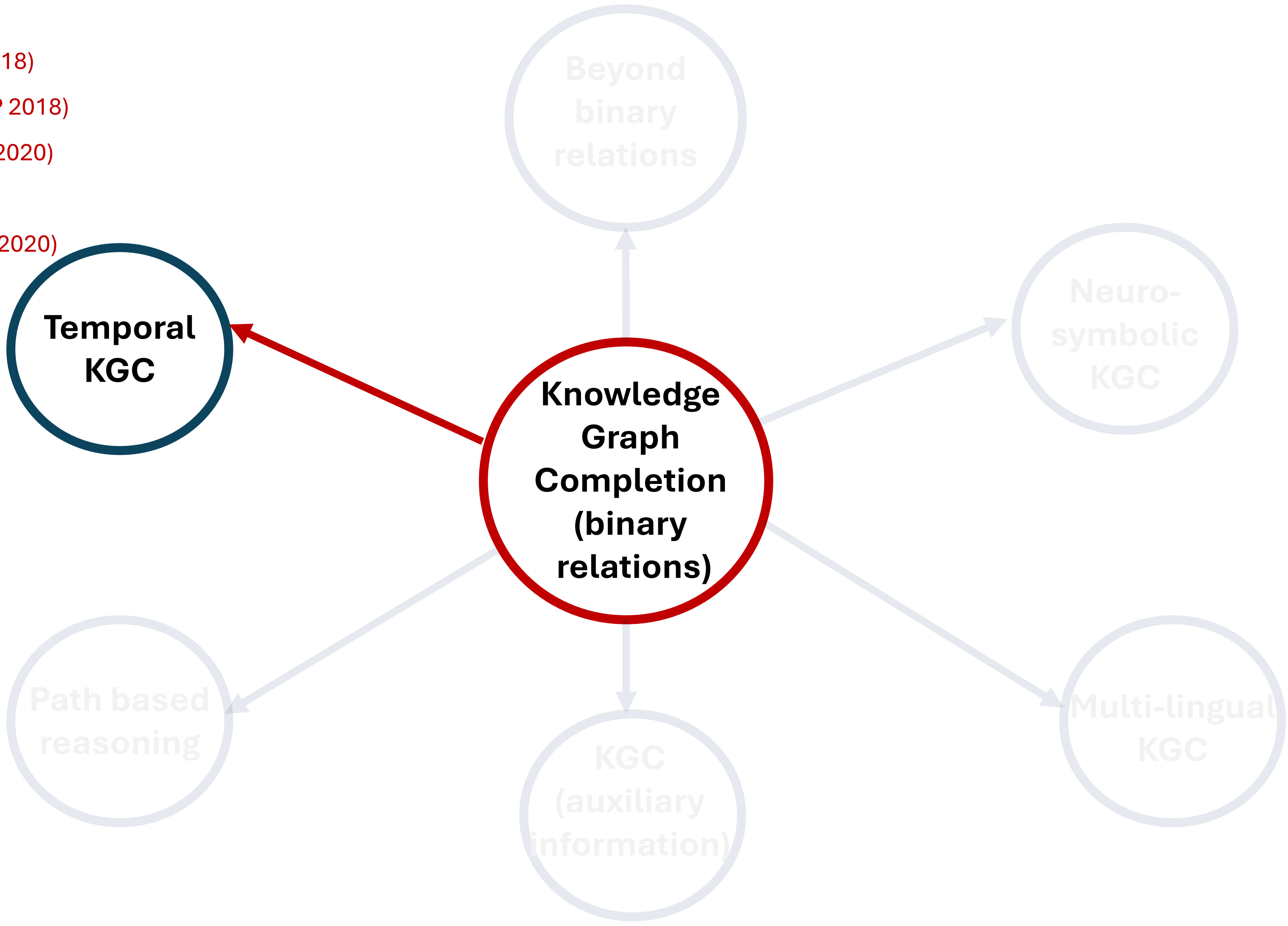
- Hypergraphs (hyper edges between nodes)
  - flies_between(British-Airways, London, New York)
  - course(Chris Manning, NLP, Fall 2024)
- New Knowledge hypergraphs were constructed:
  - FB-Auto
  - M-FB15K
- One solution to perform KGC in hypergraphs is to use convolutional neural networks (HypE (IJCAI 2020))



1. Knowledge Hypergraphs: Prediction Beyond Binary Relations, Fatemi et al, IJCAI 2020
2. Link Prediction in n-ary relational data, Gaun et al, WWW 2019
3. On the representation and embeddings of knowledge bases beyond binary relations, Wen et al, IJCAI 2016

Knowledge Graph Completion – the "spin-offs"

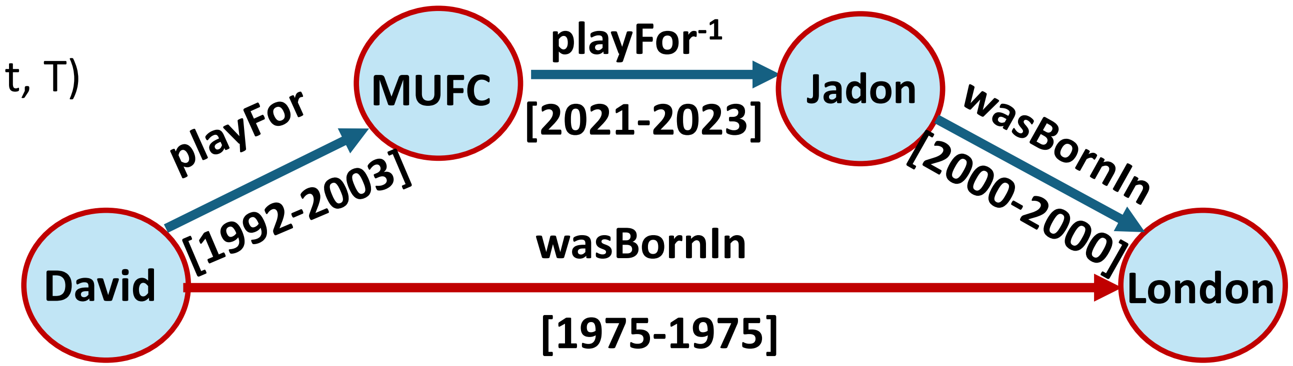
TTransE (WWW, 2018)
TA-TransE (EMNLP 2018)
DA-TransE (AAAI-2020)
HyTE (EMNLP 2018)
Timeplex (EMNLP 2020)



Temporal Knowledge Graph Completion

- Temporal KG are represented as quadruples (h, r, t, T)
- Two types of TKG:
    - T = time-instance, τ
    - T = time-interval, [t1, t2]
- Time instance Temporal Knowledge Graph:
    - Example TKG: ICEWS05-15, ICEWS14[1]
    - (Barack Obama, consulted, France, 14/06/05)
- Time Interval Knowledge Graphs
    - Examples TKG: WIKIDATA12K, YAGO11k[2]
- Temporal Knowledge Graph Completion:
    - Link Prediction: (h, r, ?, T)
    - Time prediction: (h, r, t, ?)
- Example TKGC model:
    - TTransE[3]:

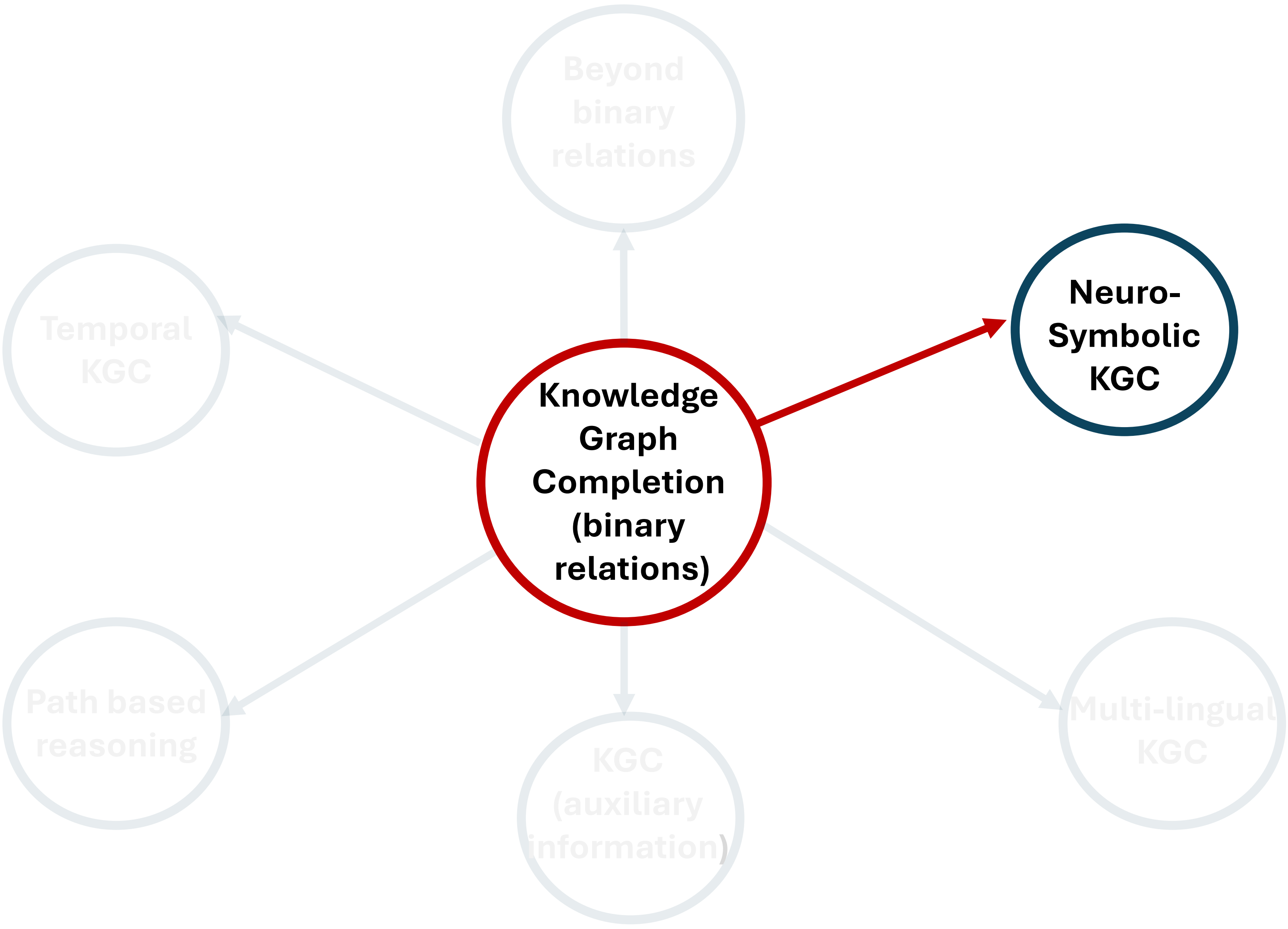
$$d(h, r, t, \tau) = ||\mathbf{h} + \mathbf{r} + \boldsymbol{\tau} - \mathbf{t}||_2^2$$



(Einstein, workedAt, ETH_Zurich, [1912,1914])

1. Learning sequence encoders for temporal knowledge graph completion, Garcıa-Duran et al, EMNLP 2018
2. HyTE: Hyperplane-based temporally aware knowledge graph embedding, DasGupta et al, EMNLP 2018
3. Deriving Validity Time in Knowledge Graph, Leblay and Chekol, WWW 2018

Knowledge Graph Completion – the "spin-offs"



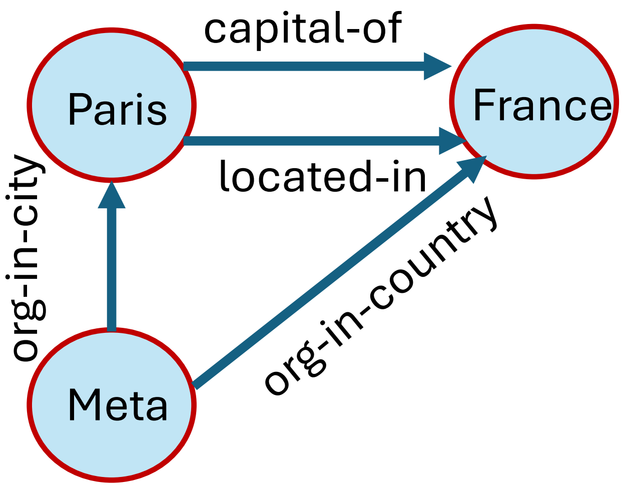
ExpressGNN **(ICLR 2020)**
RNNLogic **(ICLR 2021)**
GraphANGEL **(ICLR 2022)**
RLogic **(KDD 2022)**

Neuro-Symbolic Knowledge Graph Completion

- Utilize logical rules in addition to KG embeddings for link prediction
- $\mathcal{F}$ denotes the training rules given to you.
- We first instantiate the universally quantified (∀) rules $\mathcal{F}$ by entities in KG:
  (Paris, capital-of, France) => (Paris, Located-In, France)
- Then we aim to scoring function *I: $\mathcal{F}$ → [0,1]* to assign a truth value to each rule:
  - truth value of a rule is computed by the truth value of the constituent triples and specific connectives.
- For example, we use t-norm fuzzy logics to define the score of a rule (as composition of scores of constituents, through specific t-norm based logical connectives): [KALE model]
  - $I(f_1 \wedge f_2) = I(f_1).I(f_2)$
  - $I(f_1 \vee f_2) = I(f_1) + I(f_2) - I(f_1).I(f_2)$
  - $I(\neg f_1) = 1 - I(f_1)$
  - $I(f_1 \Rightarrow f_2) = I(f_1).I(f_2) - I(f_1) + 1$
- $\sum_{f^+ \in \mathcal{F}} \sum_{f^- \in \mathcal{N}} \left[\gamma + I(f^+) - I(f^-)\right]_+$, $f^+ \in \mathcal{F}$ positive formula, $f^- \in \mathcal{N}$ negative formula

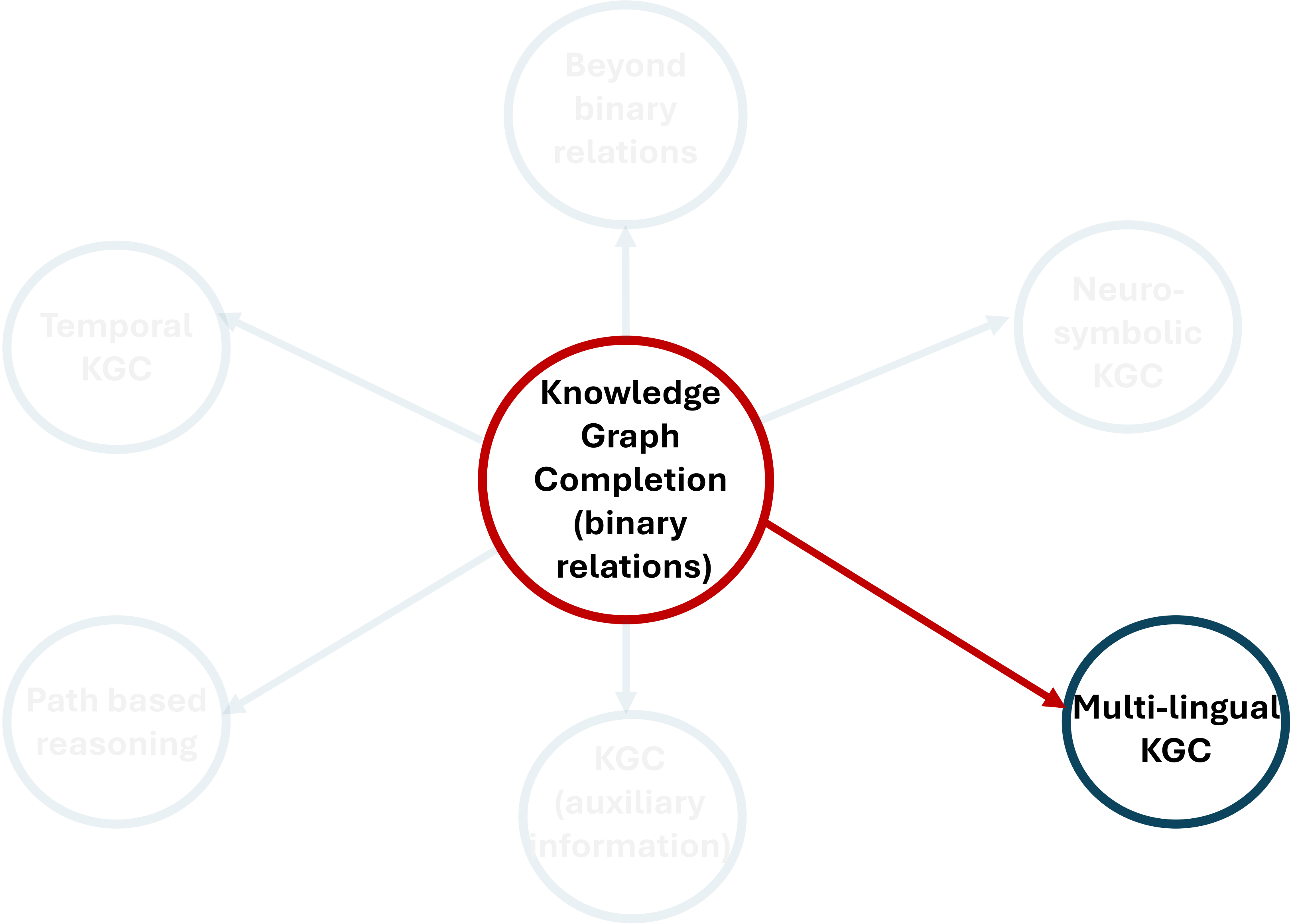
Logical Rules

$$\forall x, y \quad (x, \textit{capital-of}, y) \implies (x, \textit{located-in}, y)$$
$$\forall x, y \quad (x, \textit{org-in-city}, z) \wedge (z, \textit{located-in}, y) \implies (x, \textit{org-in-country}, y)$$
$$\forall x, y \quad (x, \textit{friend}, z) \wedge (z, \textit{hobby}, y) \implies (x, \textit{hobby}, y)$$



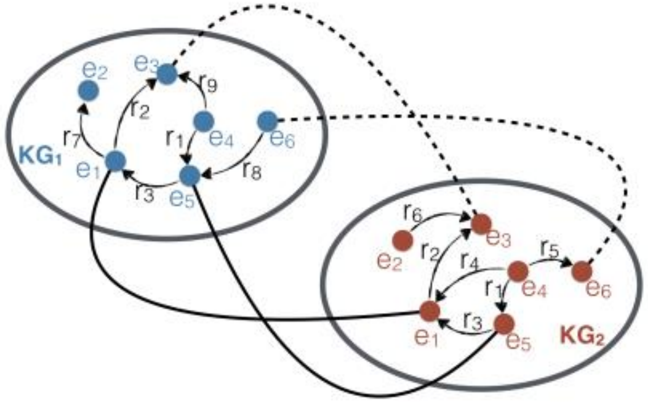
1. Jointly Embedding Knowledge Graphs and Logical Rules, Guo et al, EMNLP 2016

Knowledge Graph Completion – the “spin-offs”



MTransE **(IJCAI 2016)**
ITransE **(IJCAI 2017)**
KDCoE **(IJCAI 2018)**
BootEA **(IJCAI 2018)**
JAPE **(ISWC 2017)**

Cross-Lingual Knowledge Alignment



- Multi-lingual KGs
  - Dbpedia: English (EN), French (FR), German (De)
- Example: ((State of California, capital city, Sacremanto),
  (État de Californie, capitale, Sacramento))
- Ensure:
  - Inter-lingual Links (IIL) – two entities in two KGs are aligned
  - Triple-wise Alignment (TWA) – represents the same relations
- Goal: synchronizing different language specific versions of a knowledge graph that evolve independently
- MTransE:
  - Knowledge Model:

$$S_K = \sum_{L \in \{L_1, L_2\}} \sum_{(h,r,t) \in G_L} ||\mathbf{h} + \mathbf{r} - \mathbf{t}||$$

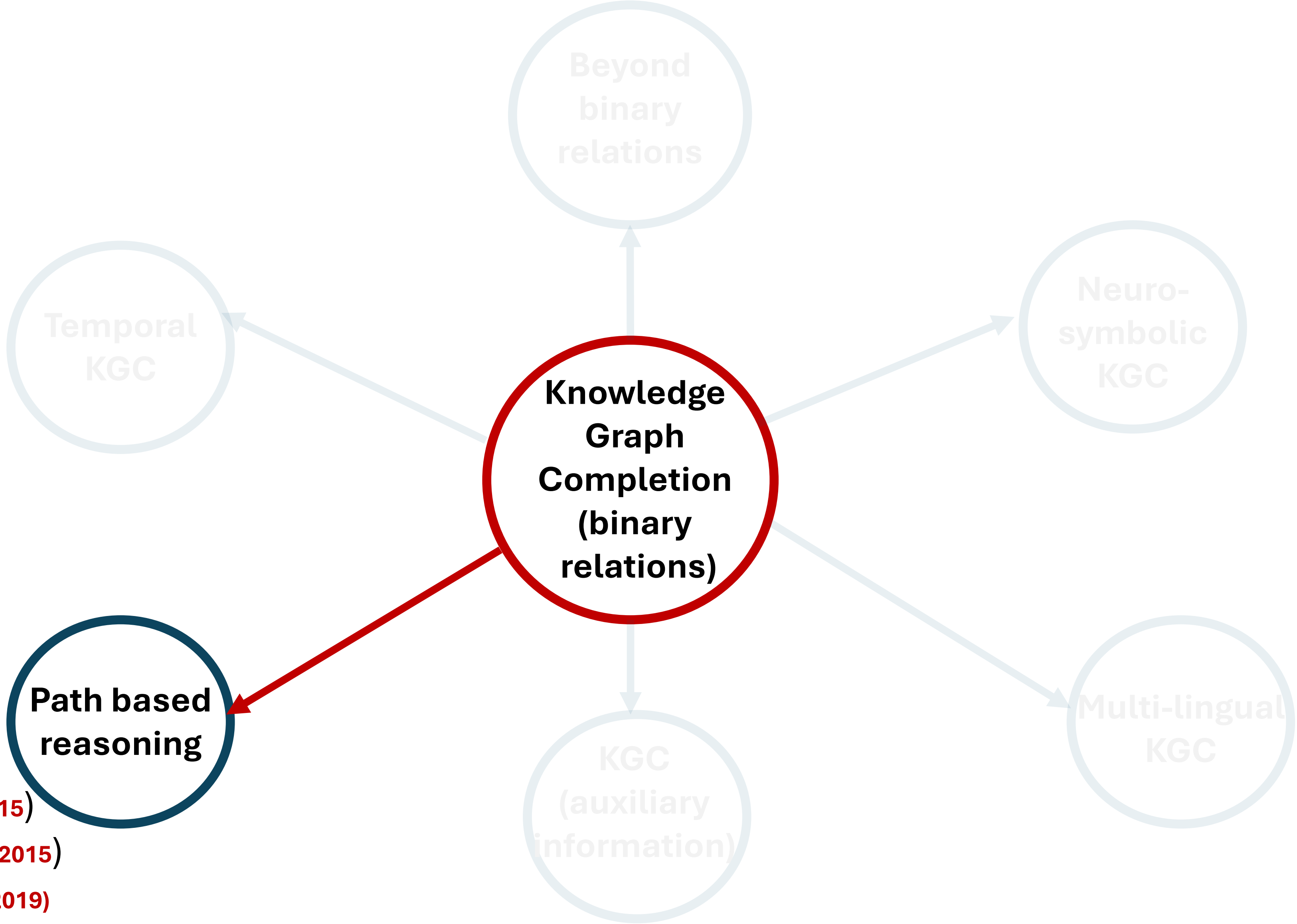
L_1, L_2: two languages ; G_L : KG in Language L

  - Alignment Model: $S_A = ||\mathbf{h}_1 - \mathbf{h}_2|| + ||\mathbf{r}_1 - \mathbf{r}_2|| + ||\mathbf{t}_1 - \mathbf{t}_2||$, $\mathbf{h}_1, \mathbf{r}_1, \mathbf{t}_1 \in G_{L_1}$; $\mathbf{h}_2, \mathbf{r}_2, \mathbf{t}_2 \in G_{L_2}$

  - Joint Training: $J = S_K + \alpha S_A$ α : hyperparameter

1. Multilingual Knowledge Graph Embeddings for Cross Lingual Knowledge Alignment, Chen et al, IJCAI 2016
2. Iterative Entity Alignment via Joint Knowledge Embeddings, Zhu et al, IJCAI 2017

Knowledge Graph Completion – the “spin-offs”



PTransE (**EMNLP 2015**)
Guu et al. (**EMNLP 2015**)
Query2Box **(ICLR 2019)**

Answering path queries

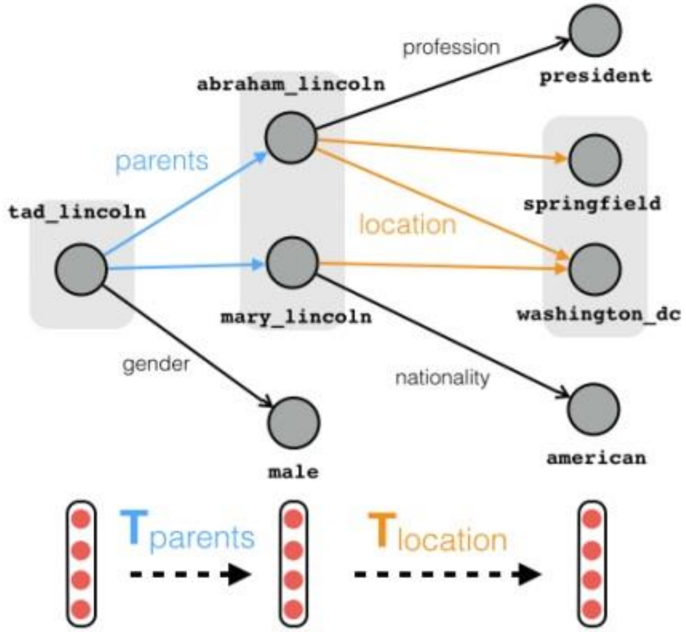
- Path queries:
  - What are the ethnicities of people from the same country as X?
  - What is the religion of X's parents?
  - What designation are there at X's institution?
  - Where did Canadian citizens with Turing Award graduate?
- KGC models can be recursively applied to answer path queries.
- Let's say we have a query
  - q: h → r_1 → r_2 →… → r_k

$$score(q,t) = ||\mathbf{h} + \mathbf{r}_1 + \ldots + \mathbf{r}_k - \mathbf{t}||$$

  - Optimization Function:

$$\sum_{t' \in N(q)} \left[1 - score(q,t) + score(q,t')\right]_+$$

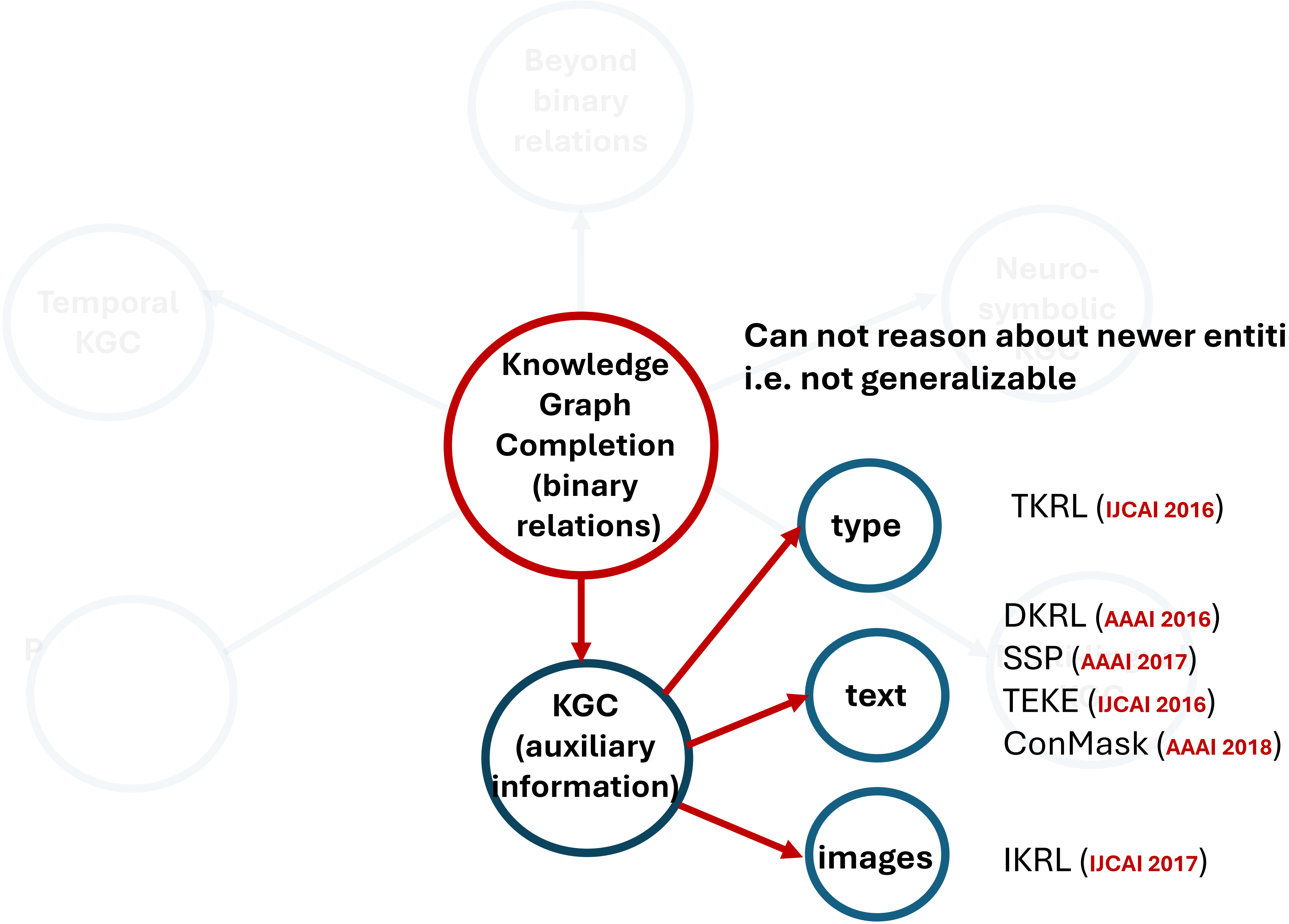
t': negative example,
t: positive example



Answering path query: *Where are Tad Lincoln's parents located ?* using low dimensional vector space.

1. Traversing Knowledge Graphs in Vector Space, Guu et al, EMNLP 2015

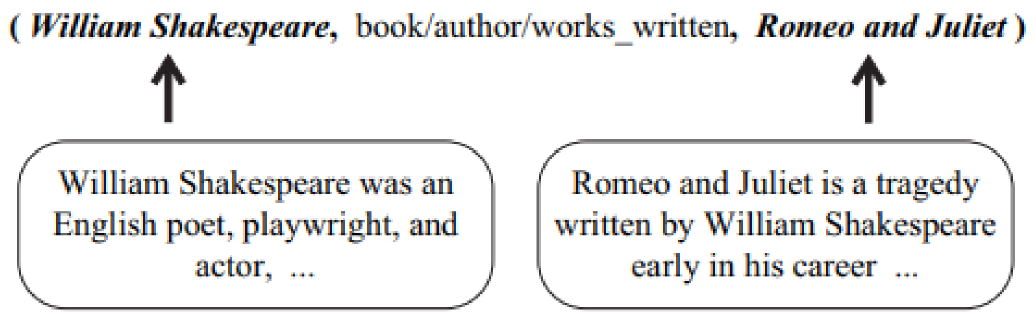
Knowledge Graph Completion – the “spin-offs”

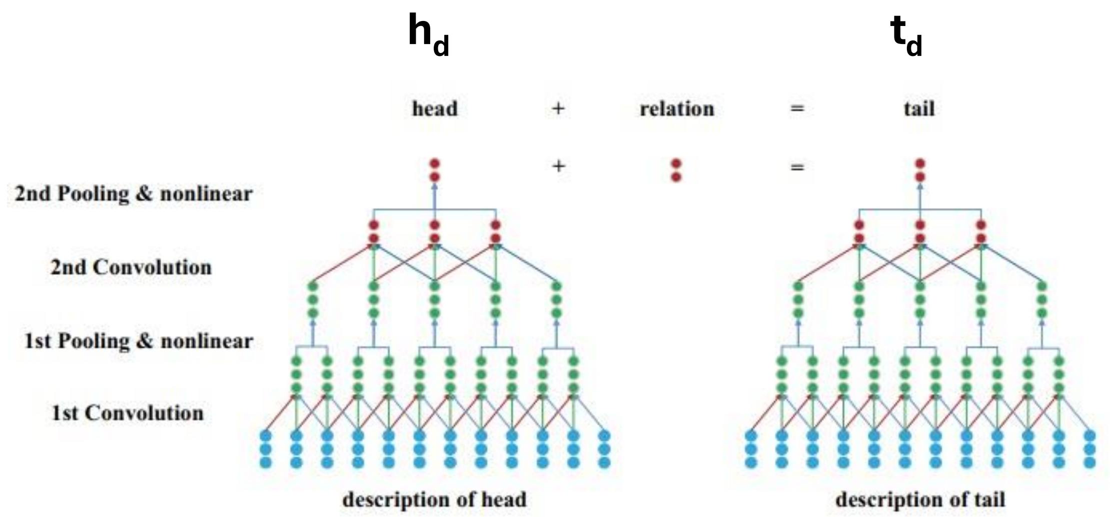


Can not reason about newer entities !
i.e. not generalizable

Knowledge Graph Completion with Entity Description

- Auxiliary Information as text description:





CNN being utilized for KGC

- For a given triple (h, r, t), we learn two type of embeddings for entities:
  - structure-based representations: $\mathbf{h_s}, \mathbf{t_s} \in \mathbb{R}^d$
  - Description-based representations: $\mathbf{h_d}, \mathbf{t_d} \in \mathbb{R}^d$
- Score S for (h, r, t):

  $S = S_S + S_D$

  $$S_S = ||\mathbf{h}_s + \mathbf{r} - \mathbf{t}_s||_2$$

- $S_D = S_{DD} + S_{DS} + S_{SD}$

$$S_{DD} = ||\mathbf{h}_d + \mathbf{r} - \mathbf{t}_d||_2 \qquad S_{DS} = ||\mathbf{h}_d + \mathbf{r} - \mathbf{t}_s||_2 \qquad S_{SD} = ||\mathbf{h}_s + \mathbf{r} - \mathbf{t}_d||_2$$

1. Representation Learning of Knowledge Graphs with Entity Description, Xie et al, AAAI 2016
2. KEPLER: A Unified Model for Knowledge Embedding and Pre-trained Language Representation (KEPLER Model) Wang et, TACL 2021

Outline

- Knowledge Graphs
- Knowledge Graph Completion
- Knowledge Graph Completion – the “spin offs”
- **Knowledge Graphs and LLMs**

Knowledge Graphs and LLMs

Interactions between Knowledge Graphs and LLMs

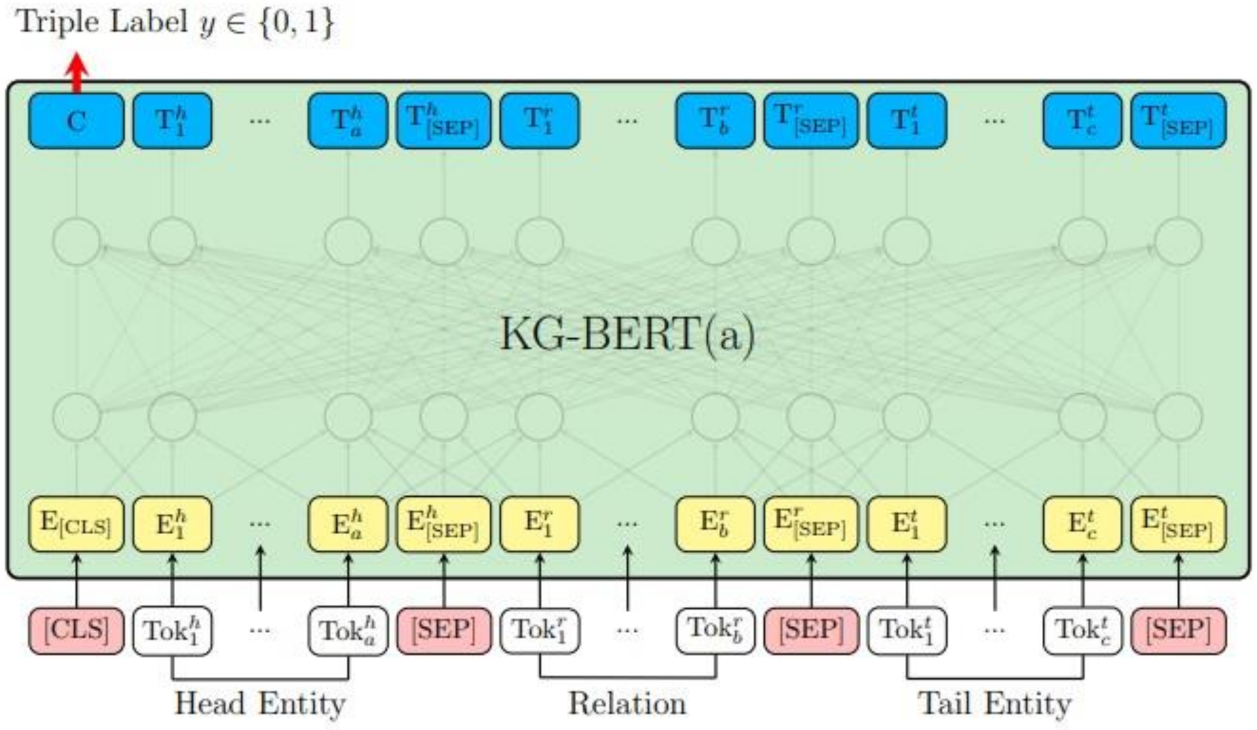
- **Using LLMs to perform KG Completion**
  - KG-BERT (AAAI 2020)
  - STAR (WWW 2021)
  - SimKGC (ACL 2022)
- **Using KGs improve LLMs**
  - ERNIE (ACL 2019)
  - KnowBERT (EMNLP 2019)
  - WKLM (ICLR 2020)
  - KEPLER (TACL 2021)
  - K-BERT (AAAI 2020)
- **Extract KGs out of LLMs**
  - LAMA (EMNLP 2019)
  - COMET (ACL 2019)
  - COMET-DISTILL (NAACL 2022)

BERT for Knowledge Graph Completion

- Finetune BERT model on the textual description of entities and relations.
- (Steve Jobs, founder, Apple Inc)
- Steve Jobs: "*Steve Jobs was an American business magnate, entrepreneur and investor.*"
- Apple Inc.: "*Apple Inc. is an American multinational company headquartered in California*"
- Finetune, by formulation a classification task:

$$s_\tau = f(h,r,t) = sigmoid(CW^T),$$
$$\text{where } W \in R^{2\times H}, C \in R^H$$
$$s_\tau \in R^2$$

$$\mathcal{L} = -\sum_{\tau \in D^+ \cup D^-} (y_\tau \log(s_{\tau 0}) + (1-y_\tau)\log(s_{\tau 1}))$$



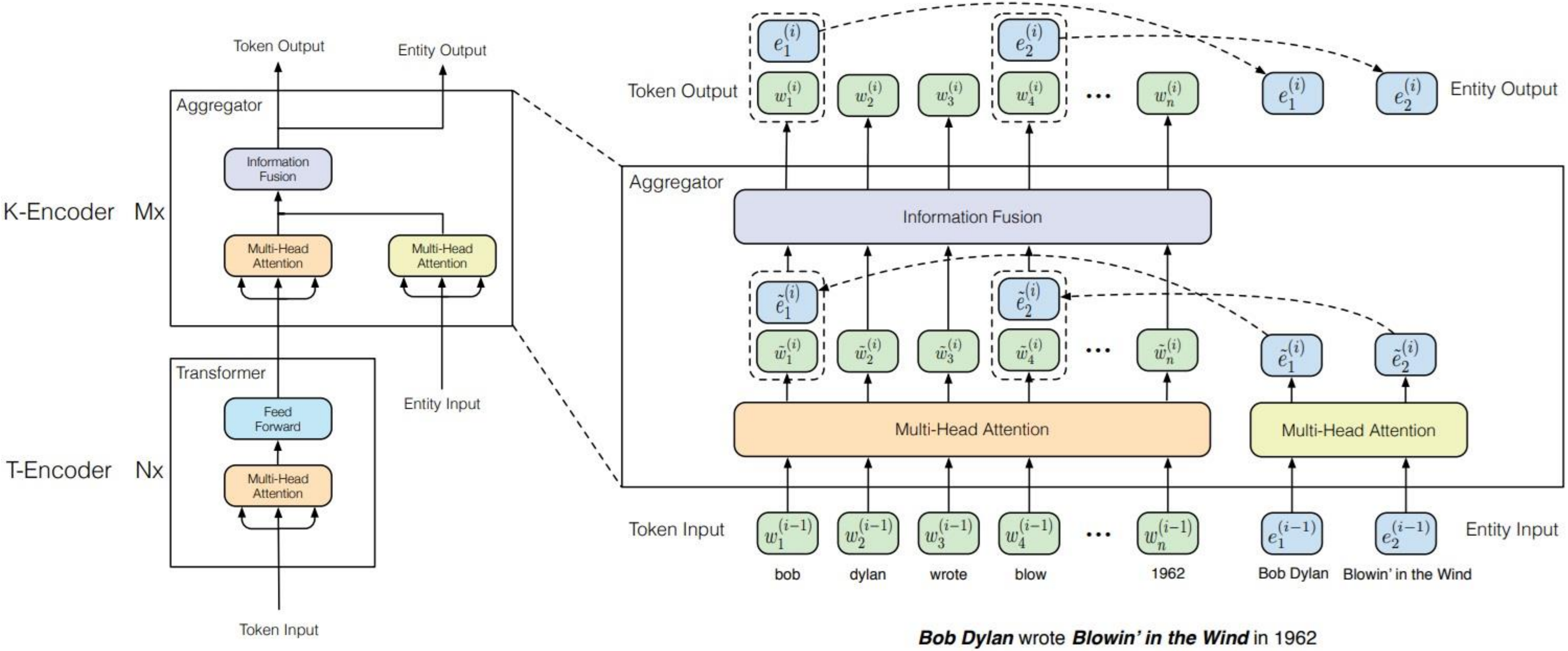
1. KGBERT: BERT for Knowledge Graph Completion, Yao et al, AAAI 2020

ERNIE: Enhanced Language Representation with Informative Entities

- They utilize both large-scale text corpora and KGs to pretrain the ERNIE model
- They recognize the named entity mentions in the training data and align these mentions to their corresponding entities in the KG
- They, then, utilize TransE model to learn the embeddings of entities mentioned in the trained text and use them as input to train the ERNIE model
- Based on the alignment between text and KG, ERNIE integrates entity representations in KG module into the layers of the semantic module

1. ERNIE: Enhanced Language Representation with Informative Entities (ERNIE Model) Zhang et al, ACL 2019

ERNIE: Enhanced Language Representation with Informative Entities



(a) Model Achitecture

(b) Aggregator

THANK YOU FOR
YOUR LİSTENİNG

DO YOU HAVE
ANY QUESTİONS?