
# To Trust or Not To Trust? Evaluation Methodology and Benchmarks for Embedding-based Knowledge Graph Completion (and beyond)

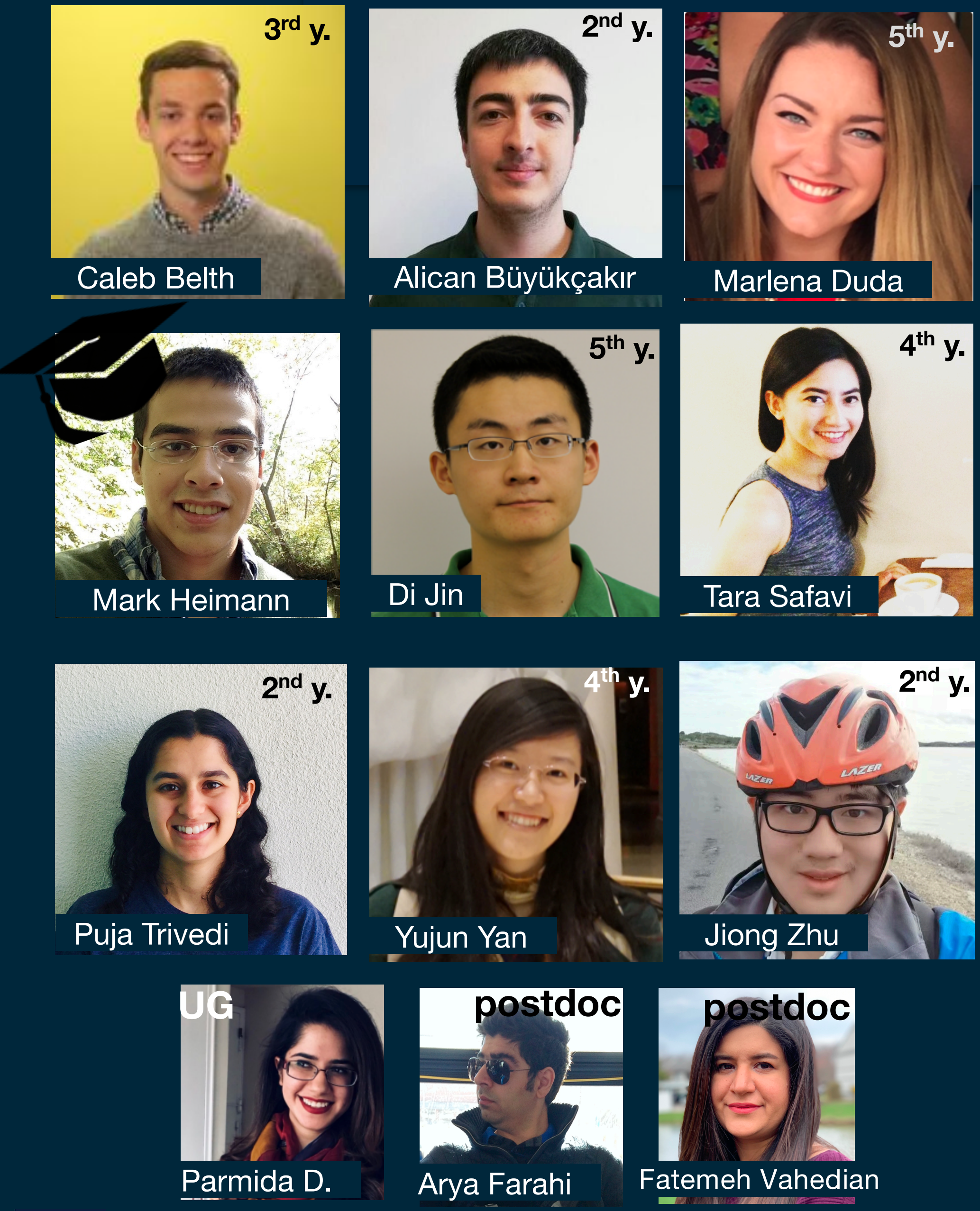
Danai Koutra
Morris Wellman Assistant Professor, CSE
Computational Medicine and Bioinformatics (courtesy)

COLING, TextGraphs workshop – December 13, 2020

*Joint work with: Caleb Belth, Edgar Meij, Tara Safavi, Jilles Vreeken, Xinyi Zheng, …*

# About me

- Danai Koutra
- Morris Wellman Assistant Professor in CSE, at the University of Michigan




GEMS Lab @ University of Michigan — Home | People | News | Research | Data and code | Lab photos

GEMS LAB

## Welcome!

We are the **Graph Exploration and Mining at Scale (GEMS)** lab at the University of Michigan, founded and led by Danai Koutra. Our team researches important data mining and machine learning problems involving interconnected data: in other words, *graphs or networks*.

From airline flights to traffic routing to neuronal interactions in the brain, graphs are ubiquitous in the real world. Their properties and complexities have long been studied in fields ranging from mathematics to the social sciences. However, many pressing problems involving graph data are still open. One well-known problem is *scalability*. With continual advances in data generation and storage capabilities, the size of graph datasets has dramatically increased, making scalable graph methods indispensable. Another is the changing nature of data. Real graphs are almost always *dynamic*, evolving over time. Finally, many important problems in the social and biological sciences involve analyzing not one but *multiple* networks.

## So, what do we do?

The problems described above call for **principled, practical, and highly scalable graph mining methods**, both theoretical and application-oriented. As such, our work connects to fields like linear algebra, distributed systems, deep learning, and even neuroscience. Some of our ongoing projects include:

- Algorithms for multi-network tasks, like matching nodes across networks
- Learning low-dimensional representations of networks in metric spaces
- Abstracting or "summarizing" a graph with a smaller network
- Analyzing network models of the brain derived from fMRI scans
- Distributed graph methods for iteratively solving linear systems
- Network-theoretical user modeling for various data science applications

We're grateful for funding from Adobe, Amazon, the Army Research Lab, the Michigan
for Data Science (MIDAS), Microsoft Azure, the National Science Foundation (NSF), an

## Interested?

If you're interested in joining our group, send an email with your interests and CV to g
opportunities@umich.edu.

COMPUTER SCIENCE & ENGINEERING UNIVERSITY OF MICHIGAN
MIDAS MICHIGAN INSTITUTE FOR DATA SCIENCE UNIVERSITY OF MICHIGAN

### News

*May 2020*
1 paper accepted to KDD'20!

*April 2020*
Caleb receives an NDSEG Fellowship!

*March 2020*
Caleb receives an NSF GRFP!

*February 2020*
Danai receives a Google Faculty Research Award!

*February 2020*
Danai was recognised as an Outstanding Senior PC Member at WSDM'20!

*January 2020*
1 paper accepted to WebConf

*January 2020*
Danai named Morris Wellman Professor!

*January 2020*
Research Fellow Fatemeh Vahedian

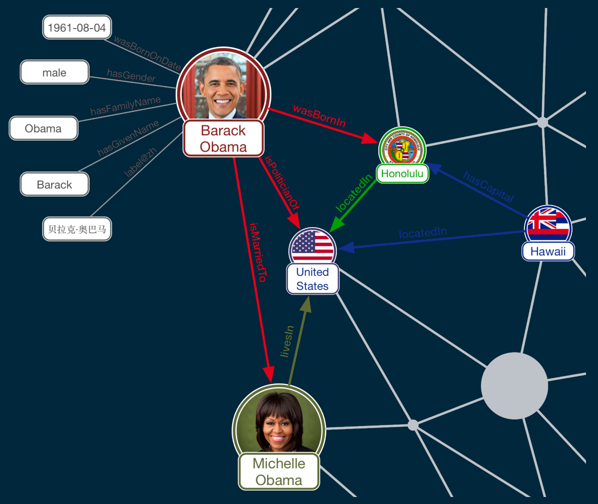
NSF
ARL
P&G
MIDAS
Adobe
amazon
Microsoft Azure

CSE GEMS LAB

++

# This talk: Knowledge Graph Completion

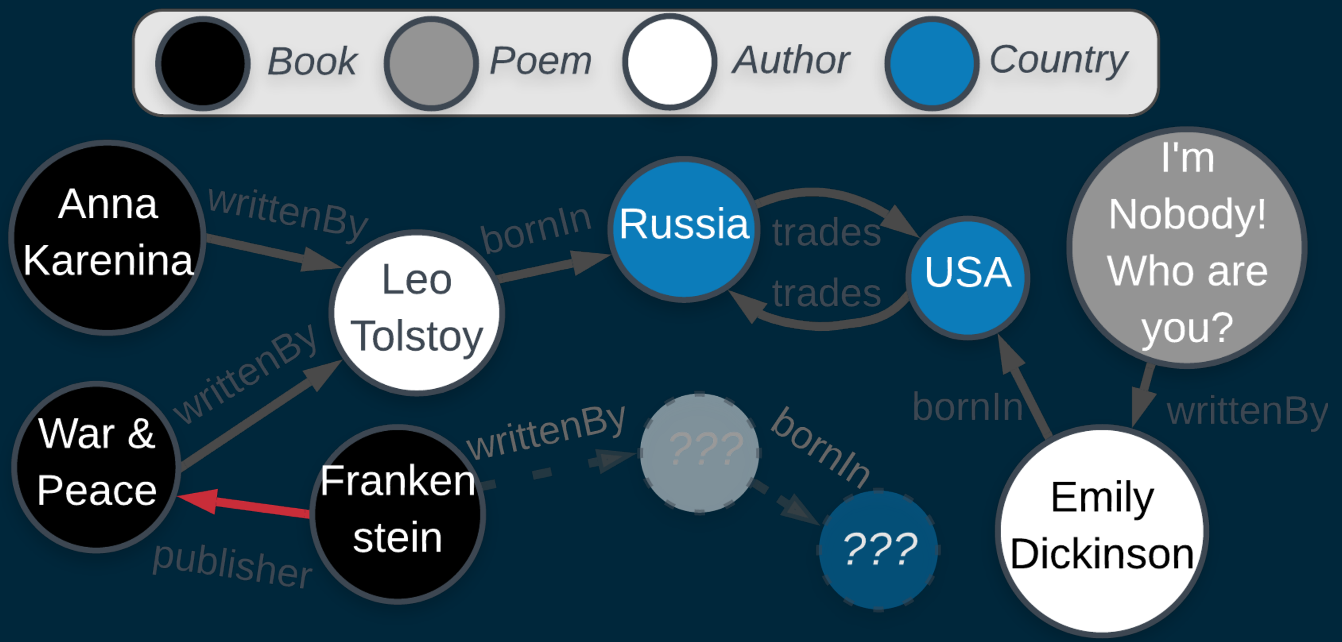
- Evaluation of knowledge graph embeddings for trustworthy link prediction [EMNLP'20a]
- CoDEx: knowledge graph completion benchmark [EMNLP'20b]
- Knowledge graph summarization for unified error detection and completion [WWW'20]




GEMS LAB

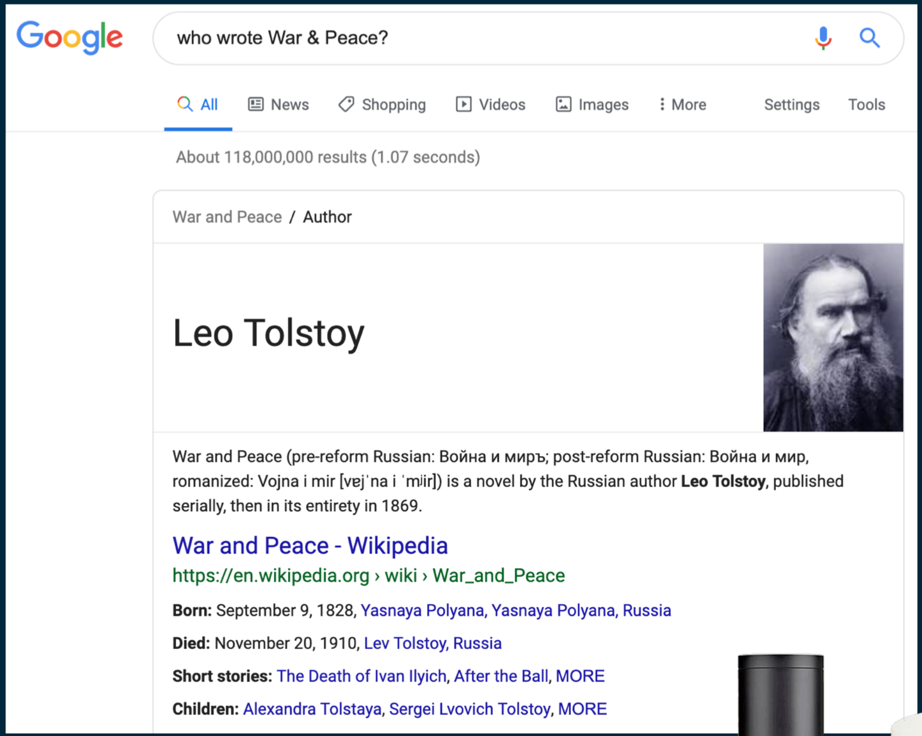
# Knowledge graphs (KGs)

store general information about the world in the structure of a graph



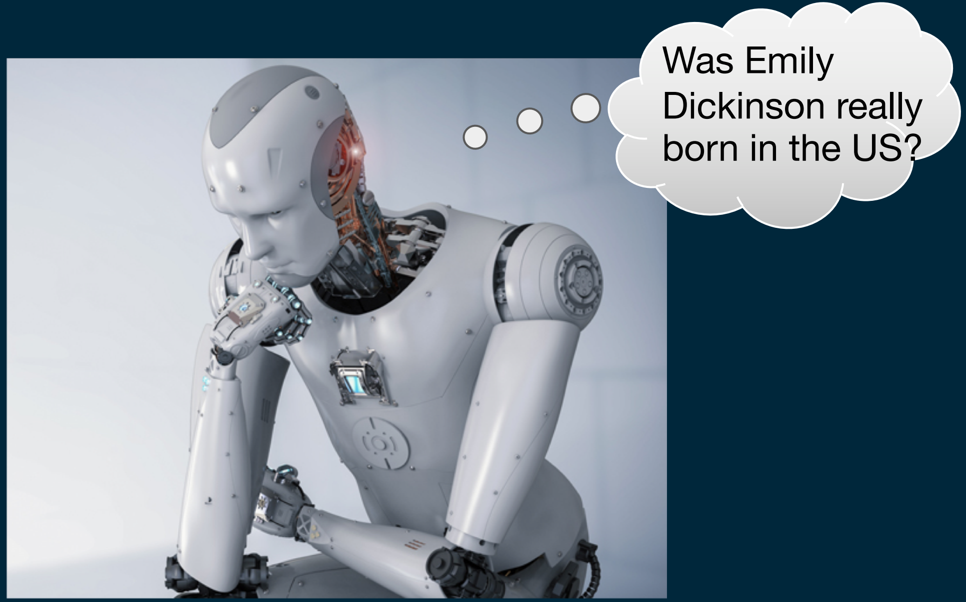
# Applications of KGs

## Question Answering





## Automatic Fact Checking



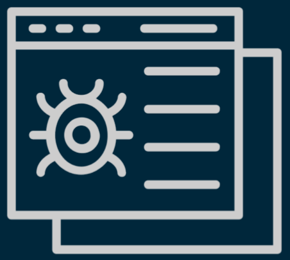

## Reading Comprehension

M CSE GEMS LAB

# KGs are constructed via

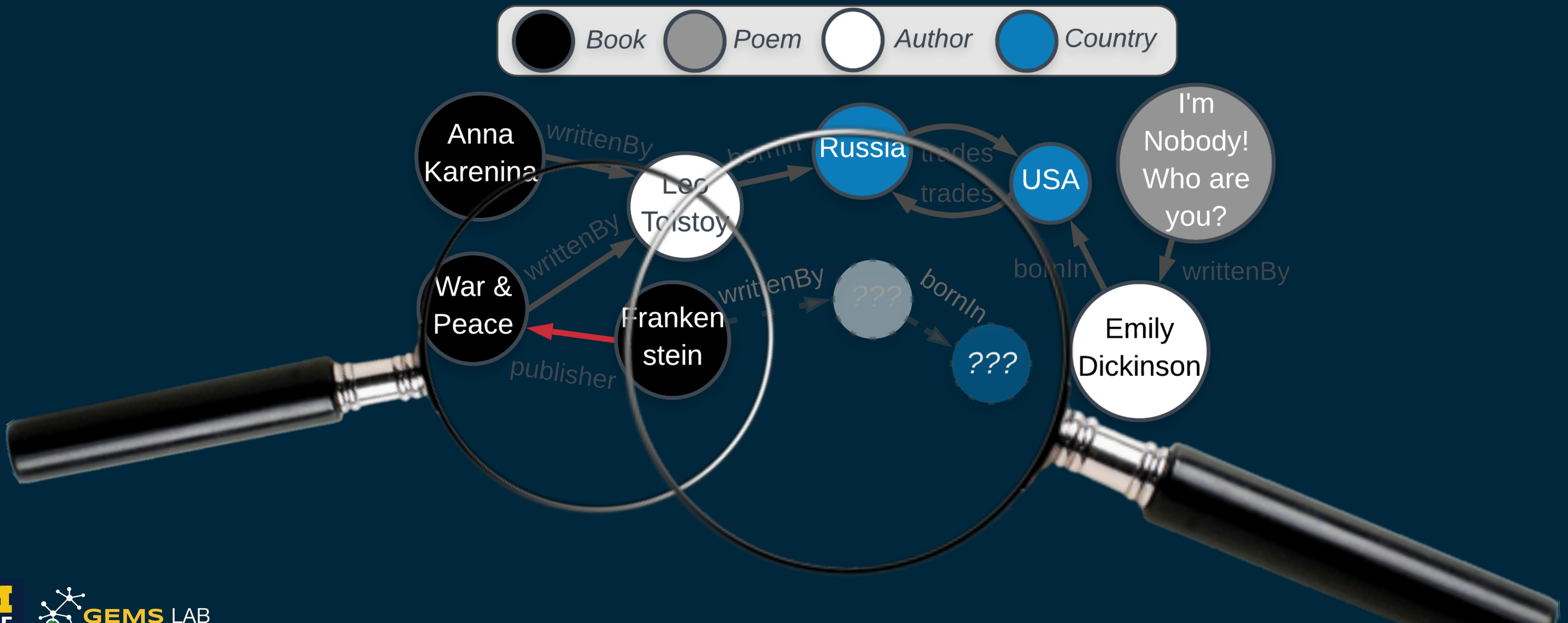
Crowd Sourcing

Web Crawling



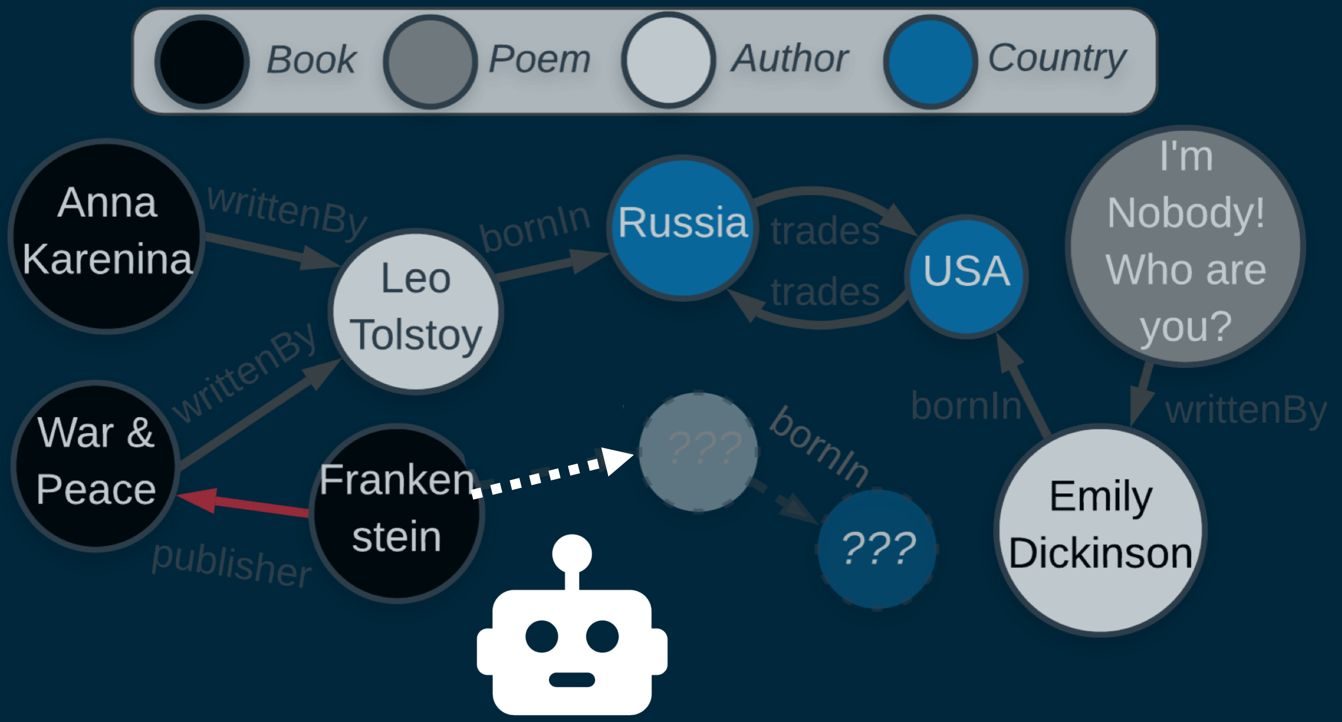
# ...which leads to

## errors and missing information



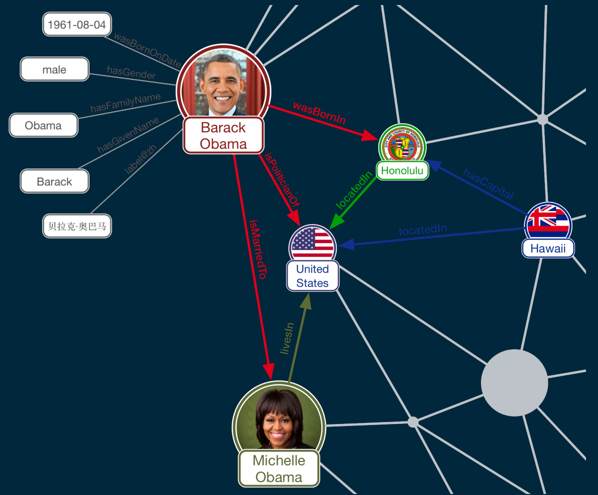
# Knowledge Graph Completion (KGC)

Automatically infer missing relationships to complete KGs



# This talk: Knowledge Graph Completion

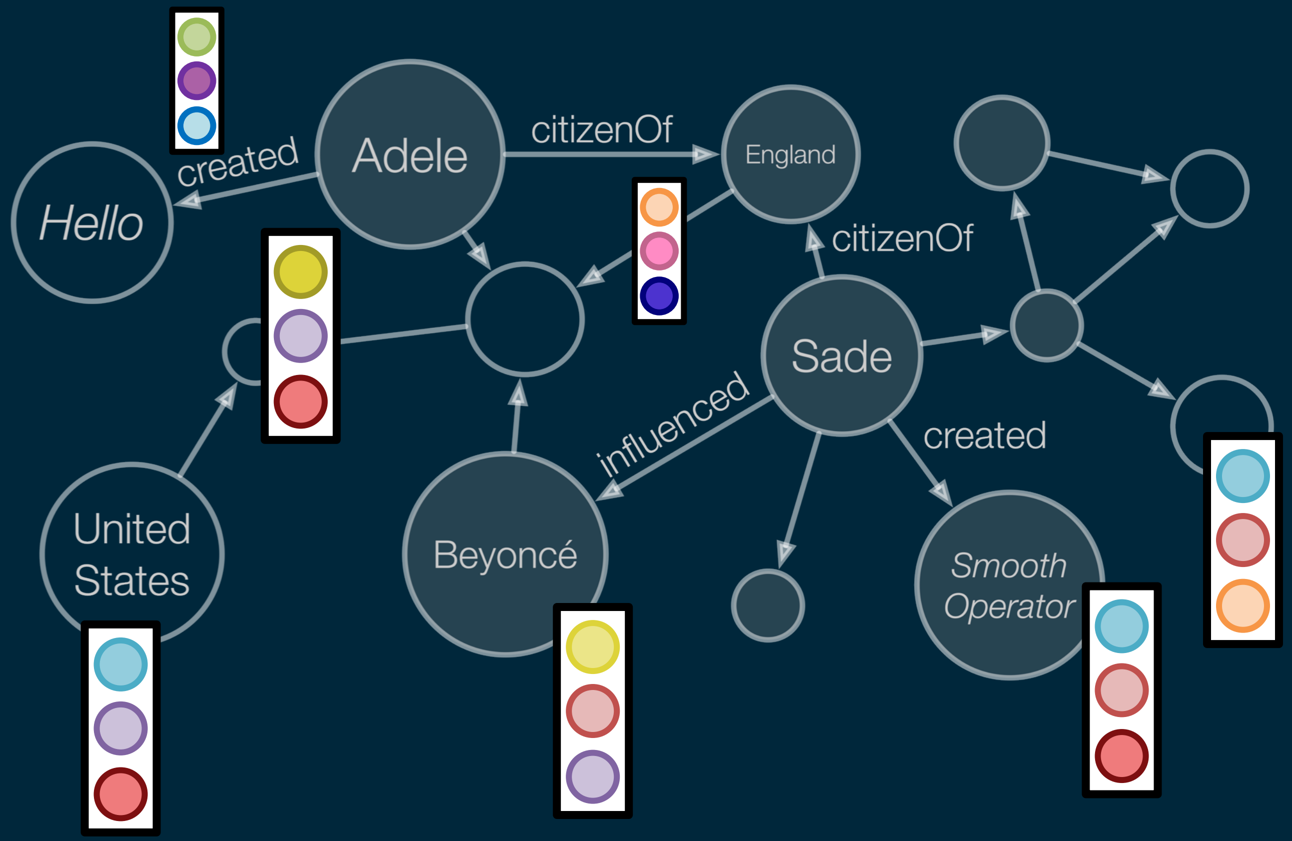
- Evaluation of knowledge graph embeddings for trustworthy link prediction [EMNLP'20a]
- CoDEx: knowledge graph completion benchmark [EMNLP'20b]
- Knowledge graph summarization for unified error detection and completion [WWW'20]




M CSE GEMS LAB

# Knowledge graph embeddings (KGE)

Latent representations of entities + relations

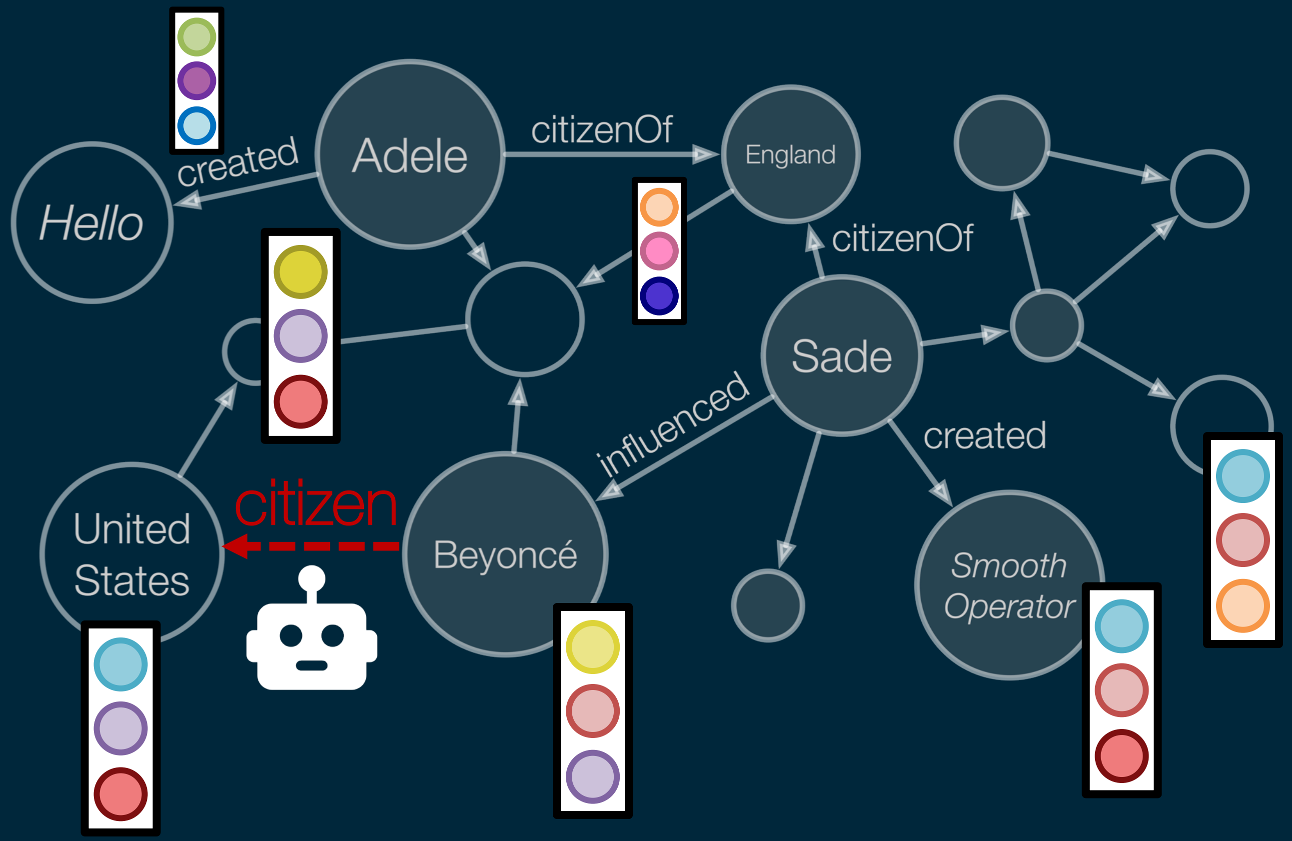





[Bordes+ NeurIPS13, Wang+ AAAI14, Yang+ ICLR15, Trouillon+ ICML16, …]

# Knowledge graph embeddings (KGE)

Used to complete KGs by predicting unseen links via ranking



# Knowledge graph embeddings (KGE)

Ranking metrics don't account for scores of predictions

Query:
(Beyoncé, citizen, ?)

| Ranked triples predicted by KGE | Uncalib. scores | True? |
|---|---|---|
| 1. (Beyoncé, citizen, India) | 0.91 | ✗ |
| 2. (Beyoncé, citizen, USA) | 0.04 | ✓ |
| 3. (Beyoncé, citizen, jazz music) | 0.02 | ✗ |
| ⋮ | ⋮ | ⋮ |

M CSE GEMS LAB

# Research question

Tara Safavi

Edgar Meij

How trustworthy are these scores?

| Ranked triples predicted by KGE | Uncalib. scores | True? |
| --- | --- | --- |
| 1. (Beyoncé, citizen, India) | 0.91 | ✗ |
| 2. (Beyoncé, citizen, USA) | 0.04 | ✓ |
| 3. (Beyoncé, citizen, jazz music) | 0.02 | ✗ |
| ⋮ | ⋮ | ⋮ |

M CSE GEMS LAB [Tara Safavi, Danai Koutra, Edgar Meij. EMNLP '20]

# Research question

In practice, prediction scores should be calibrated for deployment.

| Ranked triples predicted by KGE | Uncalib. scores | True? |
|---|---|---|
| 1. (Beyoncé, citizen, India) | 0.91 | ✗ |
| 2. (Beyoncé, citizen, USA) | 0.04 | ✓ |
| 3. (Beyoncé, citizen, jazz music) | 0.02 | ✗ |
| ⋮ | ⋮ | ⋮ |



[Dong+ KDD14, Tabacof and Costabello ICLR19, Pezeshkpour+ AKBC20]

# Contributions

### Problem

We propose to evaluate trustworthiness of KGE through the lens of calibration

### Evaluation

We investigate calibration under the closed- and open-world assumptions

### Case study

We conduct a human-AI case study to show the value of calibration

# Problem: Calibration for link prediction

| Ranked triples predicted by KGE | Calib. scores | True? |
|---|---|---|
| 1. (Beyoncé, citizen, India) | ? | ✗ |
| 2. (Beyoncé, citizen, USA) | ? | |
| 3. (Beyoncé, citizen, jazz music) | ? | |
| ⋮ | ⋮ | |

Transform scores to represent true correctness likelihoods

# Problem: Calibration for link prediction

| Ranked triples predicted by KGE | Calib. scores | True? |
|---|---|---|
| 1. (Beyoncé, citizen, India) | ? | ✗ |
| 2. (Beyoncé, citizen, USA) | ? | |
| 3. (Beyoncé, citizen, jazz music) | ? | |
| ⋮ | ⋮ | |

Prediction prob. 0.9 → 90% of predictions expected to be correct in the long run wrt a link prediction metric

# Problem: Calibration for link prediction

| Ranked triples predicted by KGE | Calib. scores | True? |
|---|---|---|
| 1. (Beyoncé, citizen, India) | ? | ✗ |
| 2. (Beyoncé, citizen, USA) | ? | |
| 3. (Beyoncé, citizen, jazz music) | ? | |
| ⋮ | ⋮ | |

Compare one-versus-all
(Platt scaling, isotonic regression)
and
multiclass (vector/matrix scaling)



[Platt ALMC99, Zadrozny and Elkan KDD02, Guo+ ICML16]

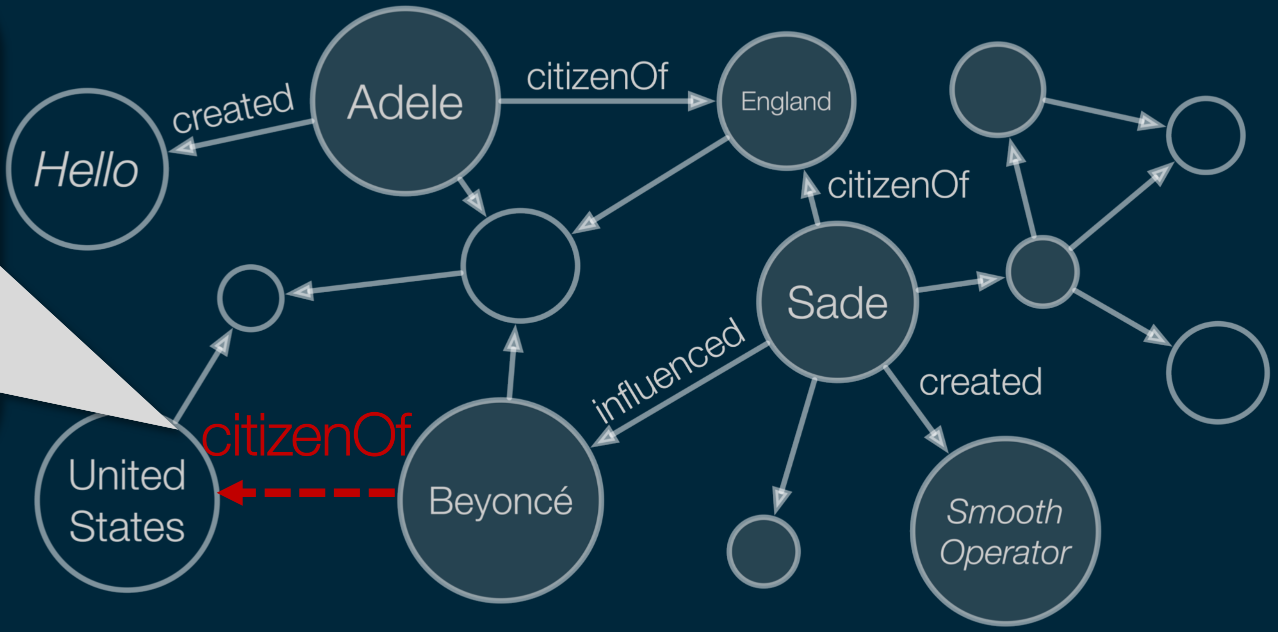
# Problem: Calibration for link prediction

| Ranked triples predicted by KGE | Calib. scores | True? |
|---|---|---|
| 1. (Beyoncé, citizen, India)<br>2. (Beyoncé, citizen, USA)<br>3. (Beyoncé, citizen, jazz music)<br>⋮ | ?<br>?<br>?<br>⋮ | ✗ |

To measure calibration, we need positive and negative examples…

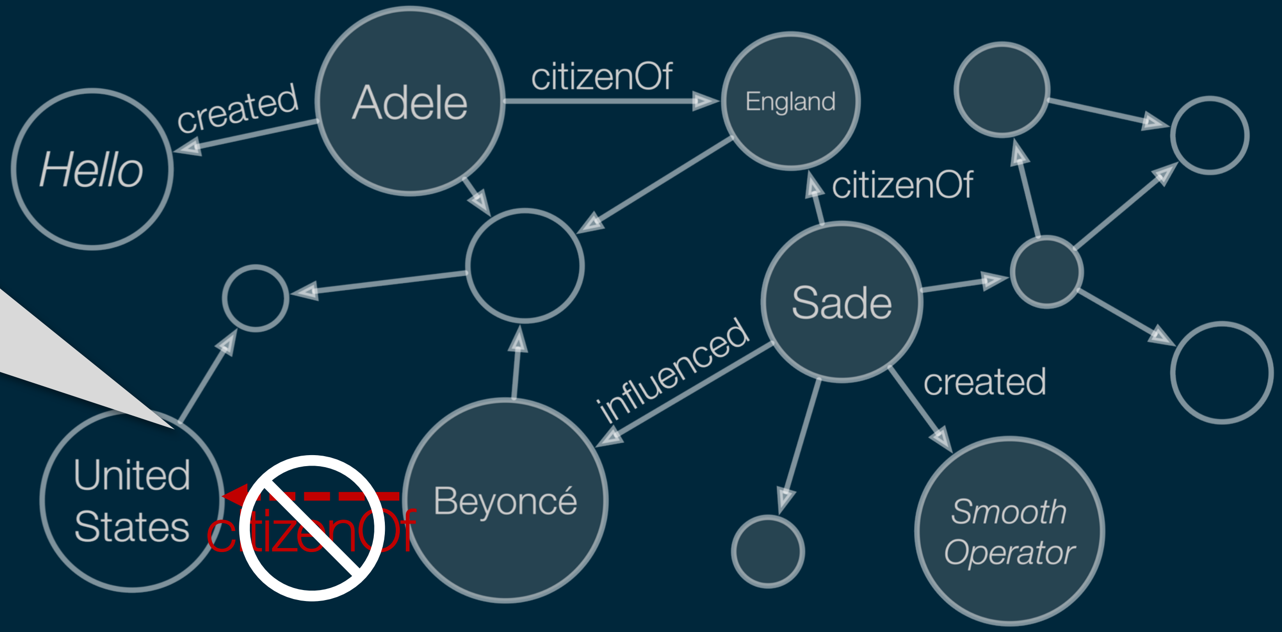
# Evaluation: Closed-world assumption (CWA)

CWA: Unseen edges considered false, measure calibration only wrt known positive edges



# Evaluation: Closed-world assumption (CWA)

CWA: A limiting assumption, but an important starting point



# CWA: Before and after calibration

| | | WN18RR | | | | | FB15K-Wiki | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | Uncalib. | One-vs-all | | Multiclass | | Uncalib. | One-vs-all | | Multiclass | |
| | | | Platt | Iso. | Vector | Matrix | | Platt | Iso. | Vector | Matrix |
| ECE (↓) | TransE | | | | | | | | | | |
| | TransH | | | | | | | | | | |
| | DistMult | | | | | | | | | | |
| | ComplEx | | | | | | | | | | |
| Acc. (↑) | TransE | | | | | | | | | | |
| | TransH | | | | | | | | | | |
| | DistMult | | | | | | | | | | |
| | ComplEx | | | | | | | | | | |

# CWA: Before and after calibration

| | | WN18RR | | | | | FB15K-Wiki | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | Uncalib. | One-vs-all | | Multiclass | | Uncalib. | One-vs-all | | Multiclass | |
| | | | Platt | Iso. | Vector | Matrix | | Platt | Iso. | Vector | Matrix |
| ECE (↓) | TransE | | | | | | | | | | |
| | TransH | | | | | | | | | | |
| | DistMult | | | | | | | | | | |
| | ComplEx | | | | | | | | | | |
| Acc. (↑) | TransE | | | | | | | | | | |
| | TransH | | | | | | | | | | |
| | DistMult | | | | | | | | | | |
| | ComplEx | | | | | | | | | | |

ECE: Expected diff in [0, 1] between average prediction prob. and (ranking) accuracy

# CWA: Before and after calibration

| | | WN18RR | | | | | FB15K-Wiki | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | Uncalib. | One-vs-all | | Multiclass | | Uncalib. | One-vs-all | | Multiclass | |
| | | | Platt | Iso. | Vector | Matrix | | Platt | Iso. | Vector | Matrix |
| ECE (↓) | TransE | 0.624 | 0.054 | 0.040 | **0.014** | 0.022 | 0.795 | 0.071 | **0.016** | 0.026 | 0.084 |
| | TransH | 0.054 | 0.057 | 0.044 | **0.018** | 0.027 | 0.177 | 0.081 | **0.024** | 0.031 | 0.089 |
| | DistMult | 0.046 | 0.040 | 0.029 | 0.044 | **0.014** | 0.104 | 0.095 | 0.031 | **0.018** | 0.054 |
| | ComplEx | 0.028 | 0.041 | 0.034 | 0.035 | **0.020** | 0.055 | 0.102 | 0.037 | **0.024** | 0.112 |
| Acc. (↑) | TransE | | | | | | | | | | |
| | TransH | | | | | | | | | | |
| | DistMult | | | | | | | | | | |
| | ComplEx | | | | | | | | | | |

Standard techniques significantly reduce error regardless of model type…

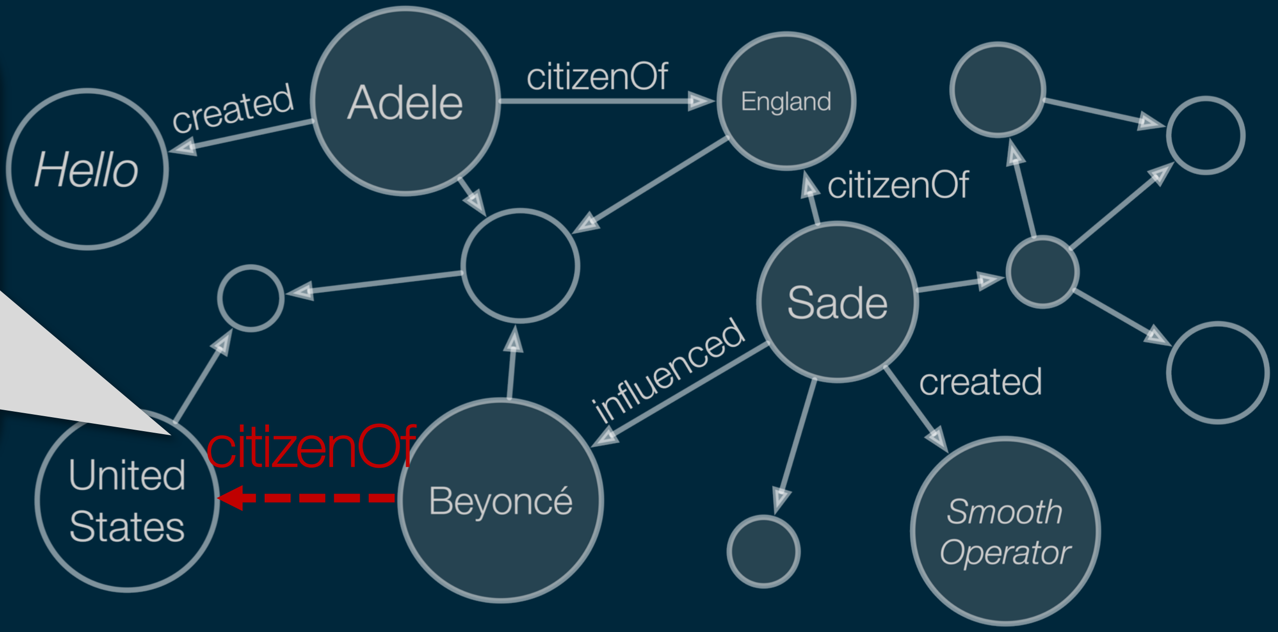
# CWA: Before and after calibration

| | | WN18RR | | | | | FB15K-Wiki | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | Uncalib. | One-vs-all | | Multiclass | | Uncalib. | One-vs-all | | Multiclass | |
| | | | Platt | Iso. | Vector | Matrix | | Platt | Iso. | Vector | Matrix |
| ECE (↓) | TransE | 0.624 | 0.054 | 0.040 | **0.014** | 0.022 | 0.795 | 0.071 | **0.016** | 0.026 | 0.084 |
| | TransH | 0.054 | 0.057 | 0.044 | **0.018** | 0.027 | 0.177 | 0.081 | **0.024** | 0.031 | 0.089 |
| | DistMult | 0.046 | 0.040 | 0.029 | 0.044 | **0.014** | 0.104 | 0.095 | 0.031 | **0.018** | 0.054 |
| | ComplEx | 0.028 | 0.041 | 0.034 | 0.035 | **0.020** | 0.055 | 0.102 | 0.037 | **0.024** | 0.112 |
| Acc. (↑) | TransE | 0.609 | 0.609 | 0.609 | 0.724 | **0.739** | 0.849 | 0.849 | 0.849 | **0.857** | 0.842 |
| | TransH | 0.625 | 0.625 | 0.625 | 0.735 | **0.740** | 0.850 | 0.850 | 0.850 | **0.858** | 0.839 |
| | DistMult | 0.570 | 0.570 | 0.570 | 0.723 | **0.761** | 0.819 | 0.819 | 0.819 | 0.862 | **0.871** |
| | ComplEx | 0.571 | 0.571 | 0.571 | 0.750 | **0.781** | 0.884 | 0.884 | 0.884 | **0.908** | 0.892 |

…and also improve ranking accuracy in some cases

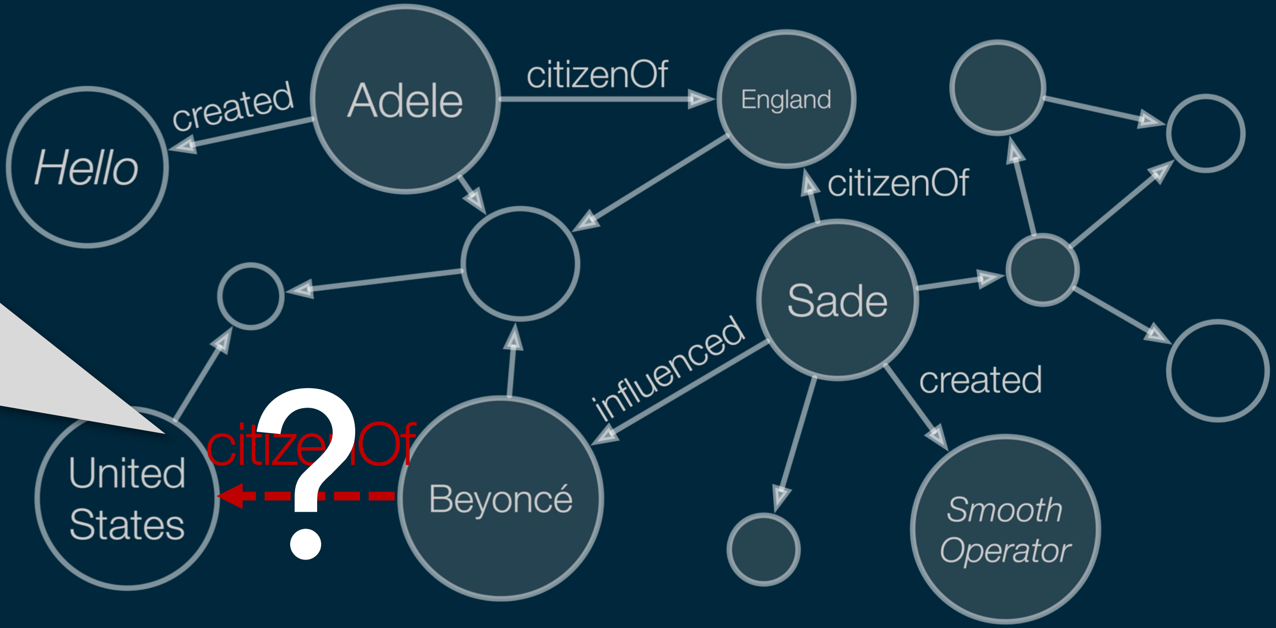
# Evaluation: Open-world assumption (OWA)

OWA: Unseen edges considered unknown until ground-truth labels are obtained



# Evaluation: Open-world assumption (OWA)

OWA: More faithful to reality, but more difficult because annotation is required



# OWA Methodology: Annotation

*The capital of the Holy Roman Empire is or was Regensburg.*

Question 1: Is this sentence factually correct? [select one]

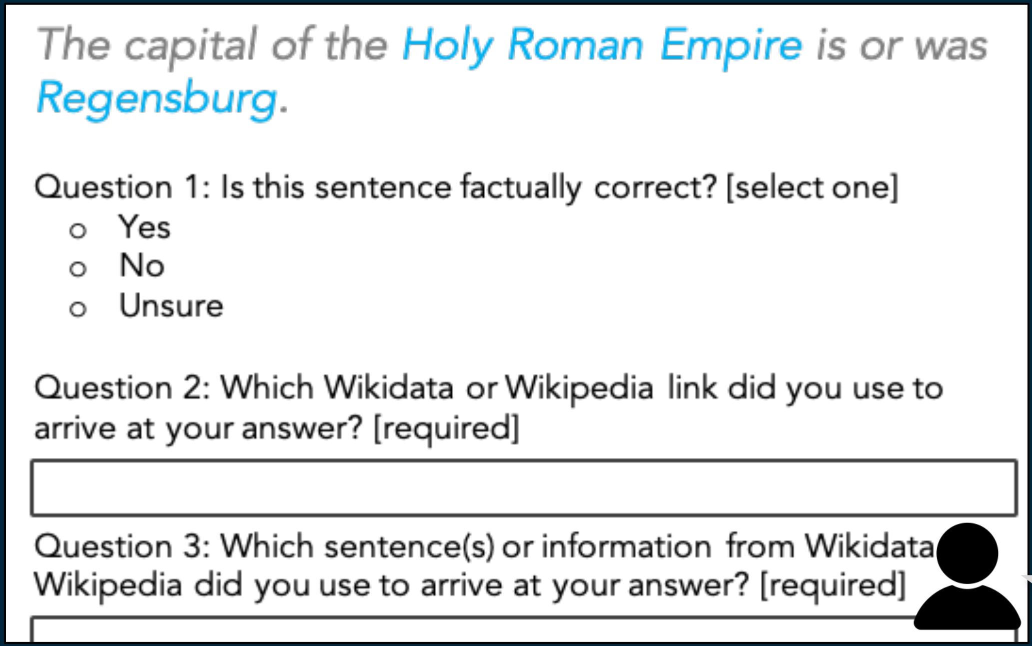
- o Yes
- o No
- o Unsure

Question 2: Which Wikidata or Wikipedia link did you use to arrive at your answer? [required]

Question 3: Which sentence(s) or information from Wikidata Wikipedia did you use to arrive at your answer? [required]

GEMS LAB [Tara Safavi, Danai Koutra, Edgar Meij. EMNLP '20]

# OWA Methodology: Annotation



*The capital of the Holy Roman Empire is or was Regensburg.*

Question 1: Is this sentence factually correct? [select one]

- Yes
- No
- Unsure

Question 2: Which Wikidata or Wikipedia link did you use to arrive at your answer? [required]

Question 3: Which sentence(s) or information from Wikidata Wikipedia did you use to arrive at your answer? [required]

Around ~1200 triples
x 5 judgments each

[Tara Safavi, Danai Koutra, Edgar Meij. EMNLP ’20]

# OWA: Before and after calibration

FB15K-237

| | ECE (↓) | | Accuracy (↑) | |
|---|---|---|---|---|
| | Uncalib. | Vector | Uncalib. | Vector |
| TransE | | | | |
| TransH | | | | |
| DistMult | | | | |
| ComplEx | | | | |
| Aggregate | | | | |

# OWA: Before and after calibration

| | ECE (↓) | | Accuracy (↑) | |
|---|---|---|---|---|
| | Uncalib. | Vector | Uncalib. | Vector |
| TransE | - | 0.234 | | |
| TransH | - | 0.307 | | |
| DistMult | 0.618 | 0.344 | | |
| ComplEx | 0.540 | 0.291 | | |
| Aggregate | 0.548 | 0.296 | | |

Standard techniques improve calibration error, but models are still too overconfident.

# OWA: Before and after calibration

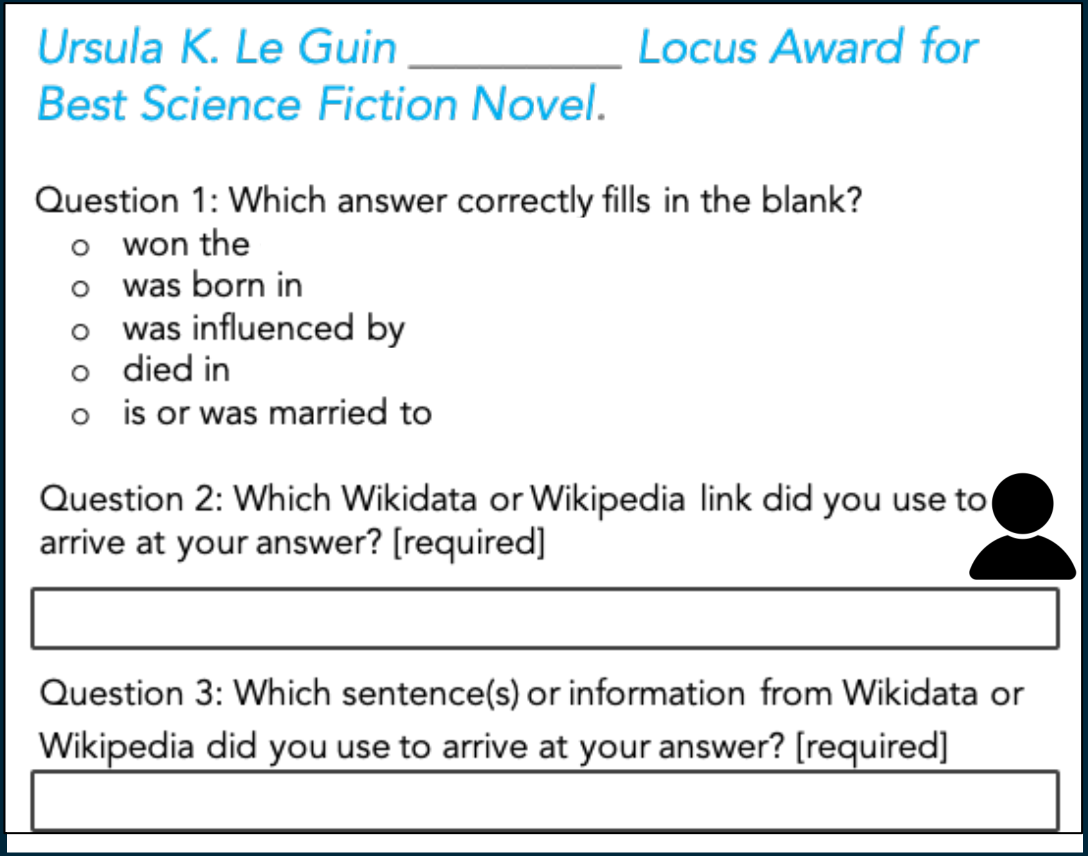
| | ECE (↓) | | Accuracy (↑) | |
|---|---|---|---|---|
| | Uncalib. | Vector | Uncalib. | Vector |
| TransE | - | 0.234 | - | 0.594 |
| TransH | - | 0.307 | - | 0.521 |
| DistMult | 0.618 | 0.344 | 0.308 | 0.509 |
| ComplEx | 0.540 | 0.291 | 0.293 | 0.581 |
| Aggregate | 0.548 | 0.296 | 0.295 | 0.549 |

Still, accuracy improves significantly → improving trustworthiness is much harder than improving accuracy

# Human-AI case study

Motivate the utility of calibration from a "trustworthiness" perspective

# Human-AI case study


*Ursula K. Le Guin __________ Locus Award for Best Science Fiction Novel.*

Question 1: Which answer correctly fills in the blank?

- won the
- was born in
- was influenced by
- died in
- is or was married to

Question 2: Which Wikidata or Wikipedia link did you use to arrive at your answer? [required]

Question 3: Which sentence(s) or information from Wikidata or Wikipedia did you use to arrive at your answer? [required]

GEMS LAB [Tara Safavi, Danai Koutra, Edgar Meij. EMNLP '20]

# Case study: No-confidence (control) group

*Ursula K. Le Guin ________ Locus Award for Best Science Fiction Novel.*

Question 1: Which answer correctly fills in the blank?

- won the
- was born in
- was influenced by
- died in
- is or was married to

Question 2: Which Wikidata or Wikipedia link did you use to arrive at your answer? [required]

Question 3: Which sentence(s) or information from Wikidata or Wikipedia did you use to arrive at your answer? [required]

Answers generated by KGE
(226 participants)

# Case study: Confidence (treatment) group

*Ursula K. Le Guin __________ Locus Award for Best Science Fiction Novel.*

Question 1: Which answer correctly fills in the blank?

- won the (50.39% confident)
- was born in (8.19% confident)
- was influenced by (5.53% confident)
- died in (14.15% confident)
- is or was married to (8.56% confident)

Question 2: Which Wikidata or Wikipedia link did you use to arrive at your answer? [required]

Question 3: Which sentence(s) or information from Wikidata or Wikipedia did you use to arrive at your answer? [required]

Answers <u>and confidence scores</u> generated by the same model (202 participants)

# Case study: Control/Treatment groups

*Ursula K. Le Guin __________ Locus Award for Best Science Fiction Novel.*

Question 1: Which answer correctly fills in the blank?

- won the (50.39% confident)
- was born in (8.19% confident)
- was influenced by (5.53% confident)
- died in (14.15% confident)
- is or was married to (8.56% confident)

Question 2: Which Wikidata or Wikipedia link did you use to arrive at your answer? [required]

Question 3: Which sentence(s) or information from Wikidata or Wikipedia did you use to arrive at your answer? [required]

## Comparisons

Completion accuracy

Completion efficiency

# Case study: Group-wise comparison

| | Accuracy ↑ | | | Sec. per triple ↓ |
|---|---|---|---|---|
| | Overall | Per triple | Per person | |
| No-conf. | | | | |
| Conf. | | | | |
| Abs. diff. | | | | |
| Rel. diff. | | | | |

[Tara Safavi, Danai Koutra, Edgar Meij. EMNLP '20]

# Case study: Group-wise comparison

Bold: significant at $p<0.05$
Underline: significant at $p<0.01$

| | Accuracy ↑ | | | Sec. per triple ↓ |
|---|---|---|---|---|
| | Overall | Per triple | Per person | |
| No-conf. | 0.8977 | 0.8969 | 0.9120 | |
| Conf. | 0.9175* | **<u>0.9220</u>** | **<u>0.9478</u>** | |
| Abs. diff. | +0.0198 | +0.0251 | +0.0358 | |
| Rel. diff. | +2.21% | +2.79% | +3.93% | |

Accuracy improves significantly in confidence group.


[Tara Safavi, Danai Koutra, Edgar Meij. EMNLP '20]

# Case study: Group-wise comparison

Bold: significant at $p<0.05$
Underline: significant at $p<0.01$

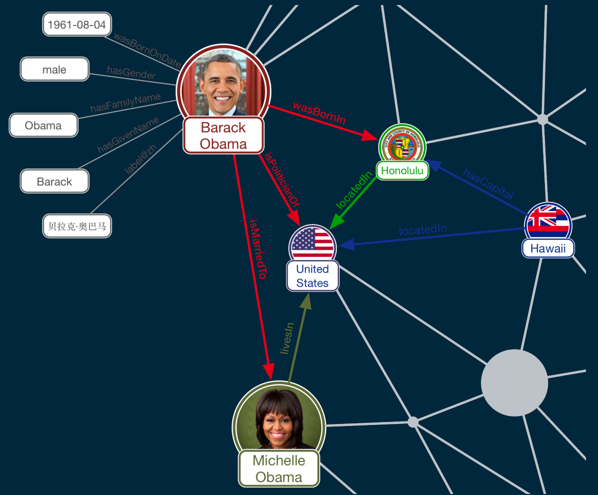
| | Accuracy ↑ | | | Sec. per triple ↓ |
|---|---|---|---|---|
| | Overall | Per triple | Per person | |
| No-conf. | 0.8977 | 0.8969 | 0.9120 | 36.88 |
| Conf. | 0.9175* | **0.9220** | **0.9478** | **31.91** |
| Abs. diff. | +0.0198 | +0.0251 | +0.0358 | -4.97 |
| Rel. diff. | +2.21% | +2.79% | +3.93% | -13.48% |

Efficiency also improves significantly in confidence group – even with quality control measures.

[Tara Safavi, Danai Koutra, Edgar Meij. EMNLP '20]

# This talk: Knowledge Graph Completion

- Evaluation of knowledge graph embeddings for trustworthy link prediction [EMNLP'20a]
- CoDEx: knowledge graph completion benchmark [EMNLP'20b]
- Knowledge graph summarization for unified error detection and completion [WWW'20]



# Forward progress requires good data

What do existing benchmarks look like in KGC?

# Most existing KGC benchmarks*

Reliance on outdated data sources

Leakage between train and test

Non-standardized versions and splits

Lack of difficult test examples

Poor interpretability



*We survey 40+ KGC papers and 12 evaluation datasets across AI/ML/NLP venues (Section 2 + Appendix A)

We survey 40+ KGC papers and 12 evaluation datasets across AI/ML/NLP venues

| | | Datasets | | | | | | | Evaluation tasks | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | Reference | FB15K | FB15K-237 | FB13 | WN18 | WN18RR | WN11 | Other | Link pred. | Triple class. | Other |
| AAAI, IJCAI | (Wang et al., 2014) | ✓ | | ✓ | ✓ | | ✓ | FB5M | ✓ | ✓ | relation extraction (FB5M) |
| | (Lin et al., 2015b) | ✓ | | ✓ | ✓ | | ✓ | FB40K | ✓ | ✓ | relation extraction (FB40K) |
| | (Wang et al., 2015) | | | | | | | NELL (Location, Sports) | ✓ | | |
| | (Nickel et al., 2016) | ✓ | | | ✓ | | | Countries | ✓ | | |
| | (Lin et al., 2016) | | | | | | | FB24K | ✓ | | |
| | (Wang and Cohen, 2016) | ✓ | | | ✓ | | | | ✓ | | |
| | (Xiao et al., 2016a) | ✓ | | ✓ | ✓ | | ✓ | | ✓ | ✓ | |
| | (Jia et al., 2016) | ✓ | | ✓ | ✓ | | ✓ | | ✓ | ✓ | |
| | (Xie et al., 2016) | ✓ | | | | | | FB15K+ | ✓ | ✓ | |
| | (Shi and Weninger, 2017) | ✓ | | | | | | SemMedDB, DBPedia | ✓ | | fact checking (not on FB15K) |
| | (Dettmers et al., 2018) | ✓ | ✓ | | ✓ | ✓ | | YAGO3-10, Countries | ✓ | | |
| | (Ebisu and Ichise, 2018) | ✓ | | | ✓ | | | | ✓ | | |
| | (Guo et al., 2018) | ✓ | | | | | | YAGO37 | ✓ | | |
| | (Zhang et al., 2020) | ✓ | ✓ | | ✓ | ✓ | | | ✓ | | |
| | (Vashishth et al., 2020a) | | ✓ | | | ✓ | | YAGO3-10 | ✓ | | |
| ICML, ICLR, NeurIPS | (Yang et al., 2015) | ✓ | | | ✓ | | | FB15K-401 | ✓ | | rule extraction (FB15K-401) |
| | (Trouillon et al., 2016) | ✓ | | | ✓ | | | | ✓ | | |
| | (Liu et al., 2017) | ✓ | | | ✓ | | | | ✓ | | |
| | (Kazemi and Poole, 2018) | ✓ | | | ✓ | | | | ✓ | | |
| | (Das et al., 2018) | | ✓ | | | ✓ | | NELL-995, UMLS, Kinship, Countries, WikiMovies | ✓ | | QA (WikiMovies) |
| | (Lacroix et al., 2018) | ✓ | ✓ | | ✓ | ✓ | | YAGO3-10 | ✓ | | |
| | (Guo et al., 2019) | ✓ | ✓ | | ✓ | | | DBPedia-YAGO3, DBPedia-Wikidata | ✓ | | entity alignment (DBPedia graphs) |
| | (Sun et al., 2019) | ✓ | ✓ | | ✓ | ✓ | | | ✓ | | |
| | (Zhang et al., 2019) | ✓ | ✓ | | ✓ | ✓ | | | ✓ | | |
| | (Balazevic et al., 2019a) | | ✓ | | | ✓ | | | ✓ | | |
| | (Vashishth et al., 2020b) | | ✓ | | | ✓ | | MUTAG, AM, PTC | ✓ | | graph classification (MUTAG, AM, PTC) |
| NAACL | (Ji et al., 2015) | ✓ | | ✓ | ✓ | | ✓ | | ✓ | ✓ | |
| | (Guo et al., 2015) | | | | | | | NELL (Location, Sports, Freq) | ✓ | ✓ | |
| | (Guu et al., 2015) | | | ✓ | | | ✓ | | ✓ | ✓ | |
| | (Garcia-Duran et al., 2015) | ✓ | | | | | | Families | ✓ | | |
| | (Lin et al., 2015a) | ✓ | | | | | | FB40K | ✓ | | relation extraction (FB40K) |
| | (Xiao et al., 2016b) | ✓ | | ✓ | ✓ | | ✓ | | ✓ | ✓ | |
| | (Nguyen et al., 2016) | ✓ | | | ✓ | | | | ✓ | | |

GEMS LAB [Tara Safavi, Danai Koutra. EMNLP '20]

# CoDEx

A set of knowledge graph **Co**mpletion **D**atasets **Ex**tracted from Wikidata and Wikipedia

CoDEx

A set of knowledge graph
Completion Datasets
Extracted from
Wikidata and Wikipedia

Well-documented, comprehensive **dataset**

**Benchmarking** in multiple KGC tasks

Comparative **case study** to set CoDEx apart

# Data Collection

Sports

Medicine

Arts

Religion

News

Science

Music

Science

Geography

Travel

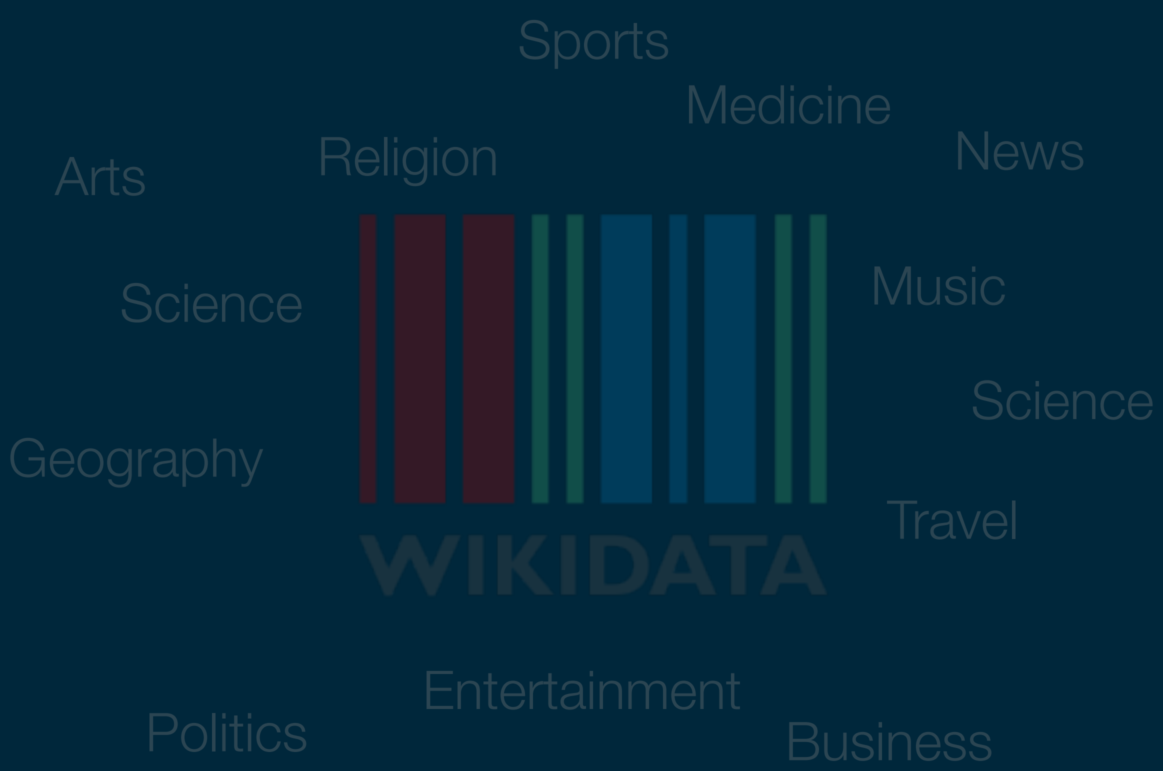
WIKIDATA

Entertainment

Politics

Business

# Data Collection




| | # entities | # relations | # triples |
|---|---|---|---|
| Codex-S | 2K | 42 | 36K |
| Codex-M | 17K | 51 | 206K |
| Codex-L | 78K | 69 | 612K |

```python
import random
codex = Codex(code="en", size="m")

eid = random.choice(list(codex.entities()))
triples = codex.triples()
triples = triples[
    (triples["head"] == eid) | (triples["tail"] == eid)
]

for (head, relation, tail) in triples.values:
    print(f"({codex.entity_label(head)},
          {codex.relation_label(relation)},
          {codex.entity_label(tail)})")
```

```
(Virginia Woolf, country of citizenship, United Kingdom)
(Virginia Woolf, occupation, diarist)
(Virginia Woolf, occupation, feminist)
(Ursula K. Le Guin, influenced by, Virginia Woolf)
(Virginia Woolf, influenced by, George Eliot)
(Virginia Woolf, genre, prose)
(Virginia Woolf, occupation, essayist)
(Leonard Sidney Woolf, spouse, Virginia Woolf)
(Virginia Woolf, genre, drama)
(Samuel R. Delany, influenced by, Virginia Woolf)
(Virginia Woolf, languages spoken, written, or signed, English)
(Gabriel García Márquez, influenced by, Virginia Woolf)
(Virginia Woolf, occupation, author)
```

# Data Collection

WIKIDATA ⟷ WIKIPEDIA

# Data Collection

```
eid = "Q51"

for code in codes:
    codex = Codex(code=code)
    print(codex.entity_label(eid))
```

```
القارة القطبية الجنوبية
Antarktika
Antarctica
Antártida
Антарктида
南极洲
```

```
codex = Codex(code="en")
print(f"From {codex.entity_wikipedia_url(eid)}:")
print(f"  '{codex.entity_extract(eid)[:400]}...'")
```

```
From https://en.wikipedia.org/wiki/Antarctica:
  'Antarctica ( or  (listen)) is Earth's southernmost contin
e, almost entirely south of the Antarctic Circle, and is sur
est continent and nearly twice the size of Australia. At 0.0
```

```
codex = Codex(code="en")
types = codex.entity_types(eid)
for etype in types:
    print(codex.entity_label(eid), "is of type", codex.entit
```

```
Antarctica is of type continent
Antarctica is of type geographic region
```

WIKIPEDIA

Entity types + text in Arabic, German, English, Spanish, Russian, Chinese

# Generating negatives for evaluation

KGs don't usually contain negatives,
which can be useful
(e.g., triple classification)

# Generating negatives for evaluation


True or false?

# Generating negatives for evaluation


True or false?

# Generating negatives for evaluation


Without realistic hard negative examples, the evaluation task is too easy!

# We generate and manually verify *hard* negatives

CoDEx

| Negative | Explanation |
|---|---|
| (*Frédéric Chopin*, *occupation*, *conductor*) | Chopin was a pianist and a composer, not a conductor. |
| (*Lesotho*, *official language*, *American English*) | English, not American English, is an official language of Lesotho. |
| (*Senegal*, *part of*, *Middle East*) | Senegal is part of West Africa. |
| (*Simone de Beauvoir*, *field of work*, *astronomy*) | Simone de Beauvoir's field of work was primarily philosophy. |
| (*Vatican City*, *member of*, *UNESCO*) | Vatican City is a UNESCO World Heritage Site but not a member state. |

# Benchmarking tasks

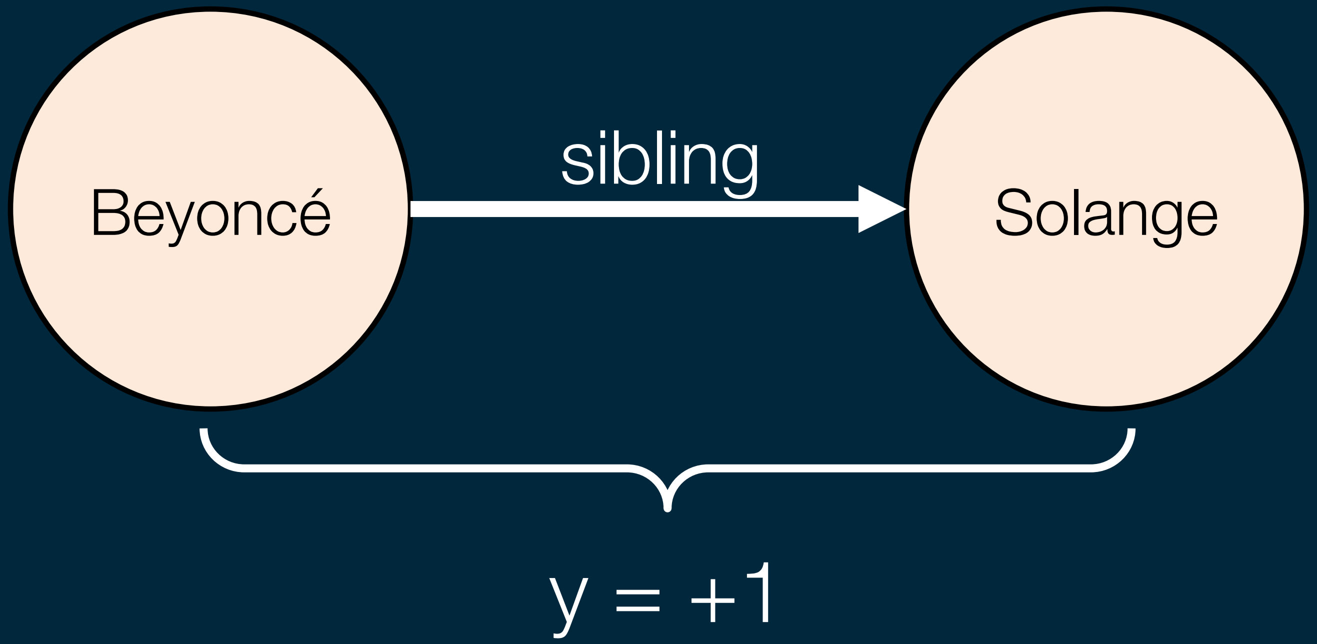
## Link prediction

Predict answers to queries like (head, relation, ?) and (?, relation, tail) by ranking candidates

# Benchmarking tasks

## Triple classification

Classify triples with labels in {-1, +1}



# Models and model selection

## Models

Linear (RESCAL, ComplEx, TuckER), translational (TransE), nonlinear (ConvE)

# Models and model selection

## Models

Linear (RESCAL, ComplEx, TuckER), translational (TransE), nonlinear (ConvE)

## Model selection

LibKGE A knowledge graph embedding library

[Ruffinelli+ ICLR20]

ADAPTIVE EXPERIMENTATION PLATFORM

# Benchmarking: Link Prediction

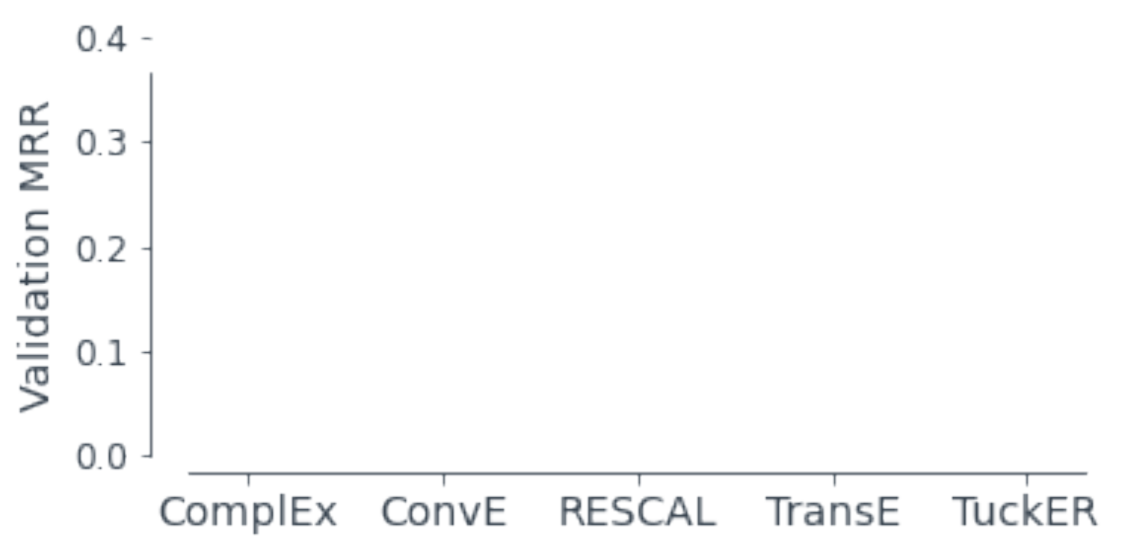
| | CoDEx-S | | | CoDEx-M | | | CoDEx-L | | |
|---|---|---|---|---|---|---|---|---|---|
| | MRR | Hits@1 | Hits@10 | MRR | Hits@1 | Hits@10 | MRR | Hits@1 | Hits@10 |
| RESCAL | | | | | | | | | |
| TransE | | | | | | | | | |
| ComplEx | | | | | | | | | |
| ConvE | | | | | | | | | |
| TuckER | | | | | | | | | |

# Benchmarking: Link Prediction

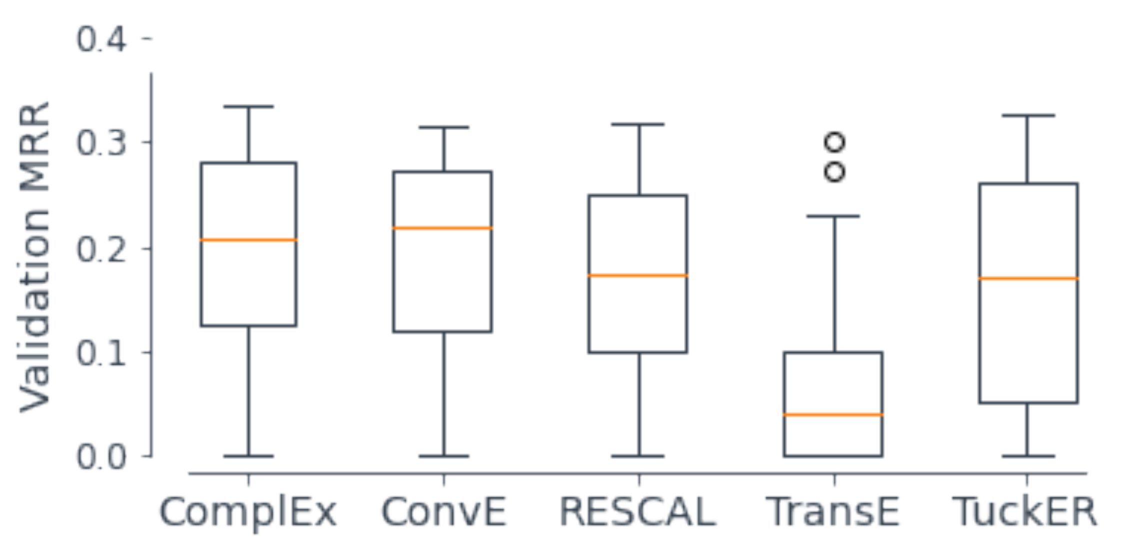
| | CoDEx-S | | | CoDEx-M | | | CoDEx-L | | |
|---|---|---|---|---|---|---|---|---|---|
| | MRR | Hits@1 | Hits@10 | MRR | Hits@1 | Hits@10 | MRR | Hits@1 | Hits@10 |
| RESCAL | 0.404 | 0.293 | 0.623 | 0.317 | 0.244 | 0.456 | 0.304 | 0.242 | 0.419 |
| TransE | 0.354 | 0.219 | 0.634 | 0.303 | 0.223 | 0.454 | 0.187 | 0.116 | 0.317 |
| ComplEx | **0.465** | **0.372** | **0.646** | **0.337** | **0.262** | **0.476** | 0.294 | 0.237 | 0.400 |
| ConvE | 0.444 | 0.343 | 0.635 | 0.318 | 0.239 | 0.464 | 0.303 | 0.240 | 0.420 |
| TuckER | 0.444 | 0.339 | 0.638 | 0.328 | 0.259 | 0.458 | **0.309** | **0.244** | **0.430** |

- Earlier models are (sometimes) stronger.
- It's important to fairly tune the models.

# Benchmarking: Link Prediction



# Benchmarking: Link Prediction




- Validation performance varies ±30% based on input configuration.
- Loss function affects performance most (best: cross-entropy).

# Benchmarking: Triple Classification

Different negative generation strategies

| | CoDEx-S | | | | | | CoDEx-M | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Uniform | | Relative freq. | | Hard neg. | | Uniform | | Relative freq. | | Hard neg. | |
| | Acc. | F1 | Acc. | F1 | Acc. | F1 | Acc. | F1 | Acc. | F1 | Acc. | F1 |
| RESCAL | | | | | | | | | | | | |
| TransE | | | | | | | | | | | | |
| ComplEx | | | | | | | | | | | | |
| ConvE | | | | | | | | | | | | |
| TuckER | | | | | | | | | | | | |

# Benchmarking: Triple Classification

Different negative generation strategies

| | CoDEx-S | | | | | | CoDEx-M | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Uniform | | Relative freq. | | Hard neg. | | Uniform | | Relative freq. | | Hard neg. | |
| | Acc. | F1 | Acc. | F1 | Acc. | F1 | Acc. | F1 | Acc. | F1 | Acc. | F1 |
| RESCAL | 0.972 | 0.972 | 0.916 | 0.920 | **0.843** | **0.852** | 0.977 | 0.976 | 0.921 | 0.922 | 0.818 | 0.815 |
| TransE | 0.974 | 0.974 | 0.919 | 0.923 | 0.829 | 0.837 | **0.986** | **0.986** | 0.932 | 0.933 | 0.797 | 0.803 |
| ComplEx | **0.975** | **0.975** | **0.927** | **0.930** | 0.836 | 0.846 | 0.984 | 0.984 | 0.930 | 0.933 | 0.824 | 0.818 |
| ConvE | 0.972 | 0.972 | 0.921 | 0.924 | 0.841 | 0.846 | 0.979 | 0.979 | **0.934** | **0.935** | **0.826** | **0.829** |
| TuckER | 0.973 | 0.973 | 0.917 | 0.920 | 0.840 | 0.846 | 0.977 | 0.977 | 0.920 | 0.922 | 0.823 | 0.816 |

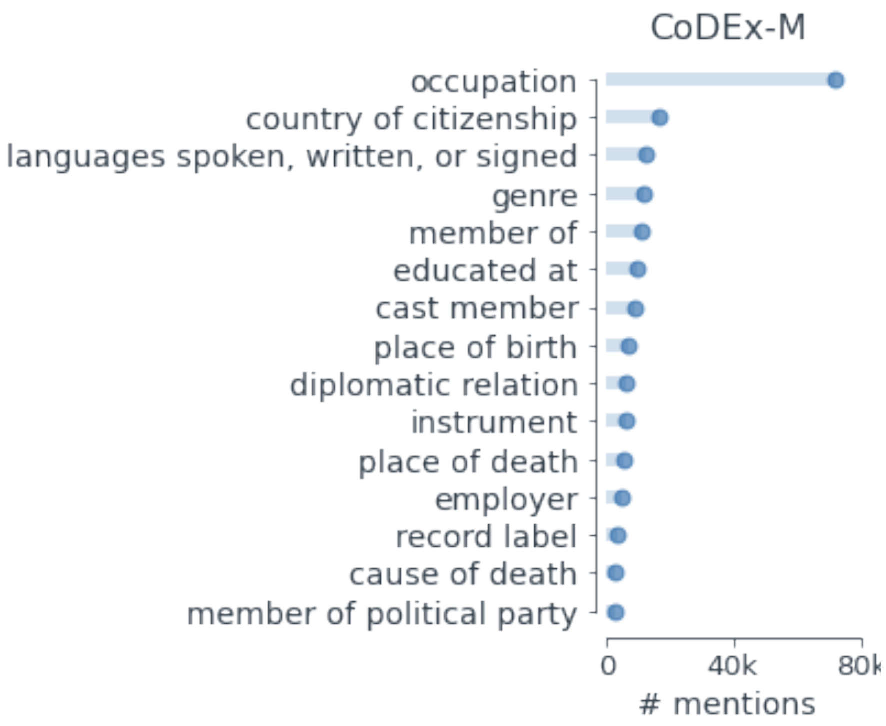
Accuracy drops up to 19 points on hard negative examples compared to randomly generated negatives.

# Comparative Analysis

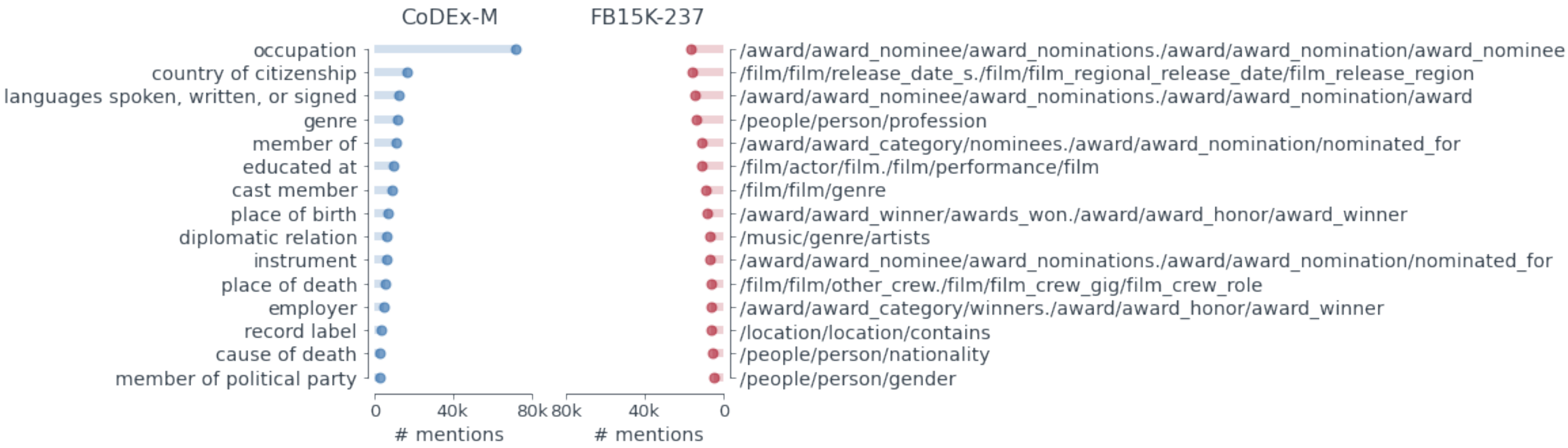
CoDEx vs Freebase

FB15K-237 [Toutanova and Chen 2015]

# Content Comparison



# Content Comparison



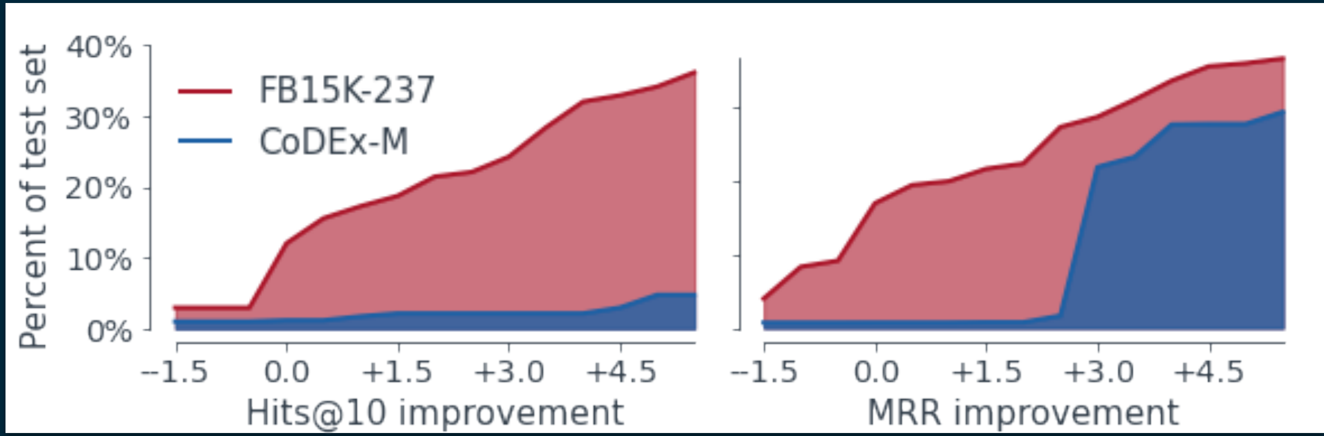

CoDEx covers a wider selection of content and is easier to interpret.

GEMS LAB [Tara Safavi, Danai Koutra. EMNLP '20] CoDEx

# Difficulty Comparison

We devise a non-learning baseline that answers link prediction queries based on entity frequency

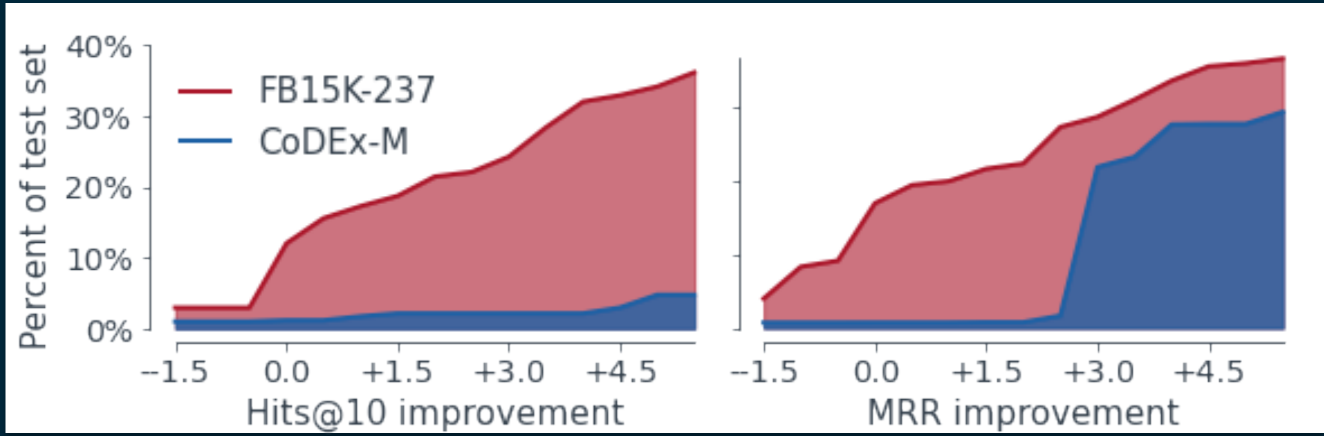
# Difficulty Comparison




## Surprisingly…

…the baseline outperforms the best model on FB15K-237 for ~10% of the dataset, and is within 5 points for ~40%!

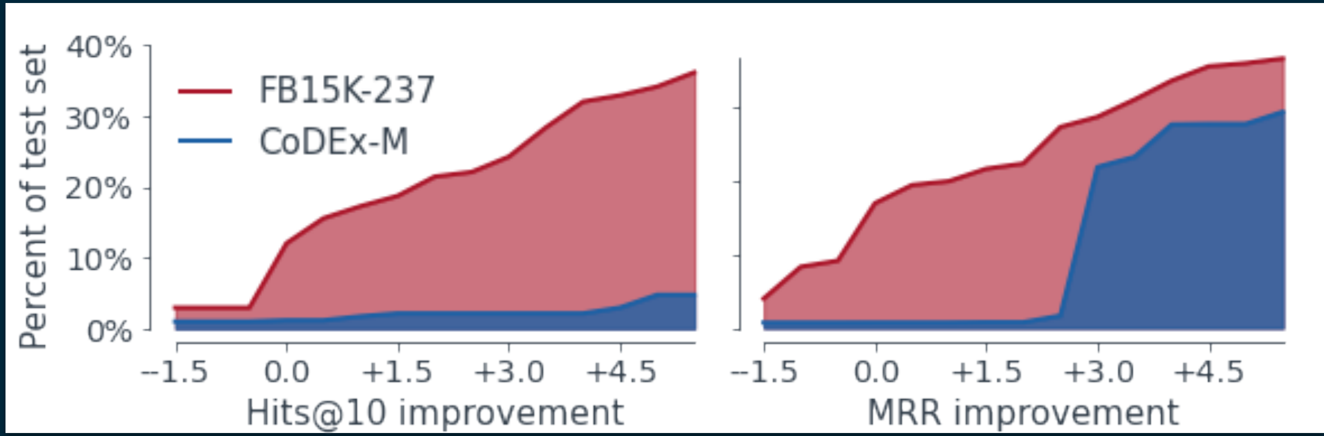
# Difficulty Comparison




## Why?

FB15K-237 is skewed toward a few entities (e.g., USA, male) and contains non-binary relations with few possible values

# Difficulty Comparison

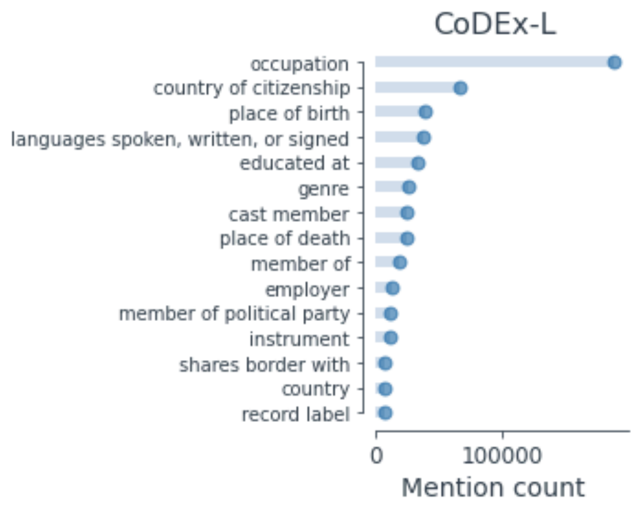



## tl;dr

FB15K-237 doesn't require as much complex reasoning as CoDEx – easier to model with just frequency patterns

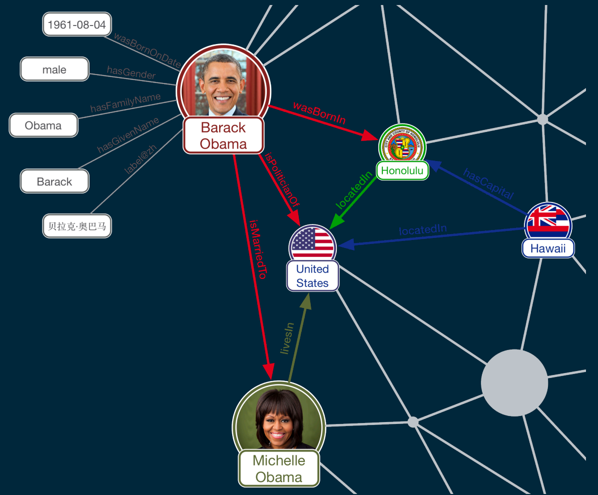
# Explore CoDEx.ipynb

```
count_df = count_relations(triples)
count_df["label"] = [
    codex.relation_label(rid) for rid in count_df["relation"]]

k = 15

ax = plot_top_k(
    count_df,
    k=k,
    color=palette[-1],
    linewidths=6,
    figsize=(5, 4)
)

ax.set_xscale("linear")
ax.set_xlabel("Mention count", fontsize=14)
ax.set_title(codex.name(), fontsize=16)
ax.tick_params("x", labelsize=12)

plt.tight_layout()
plt.show()
```



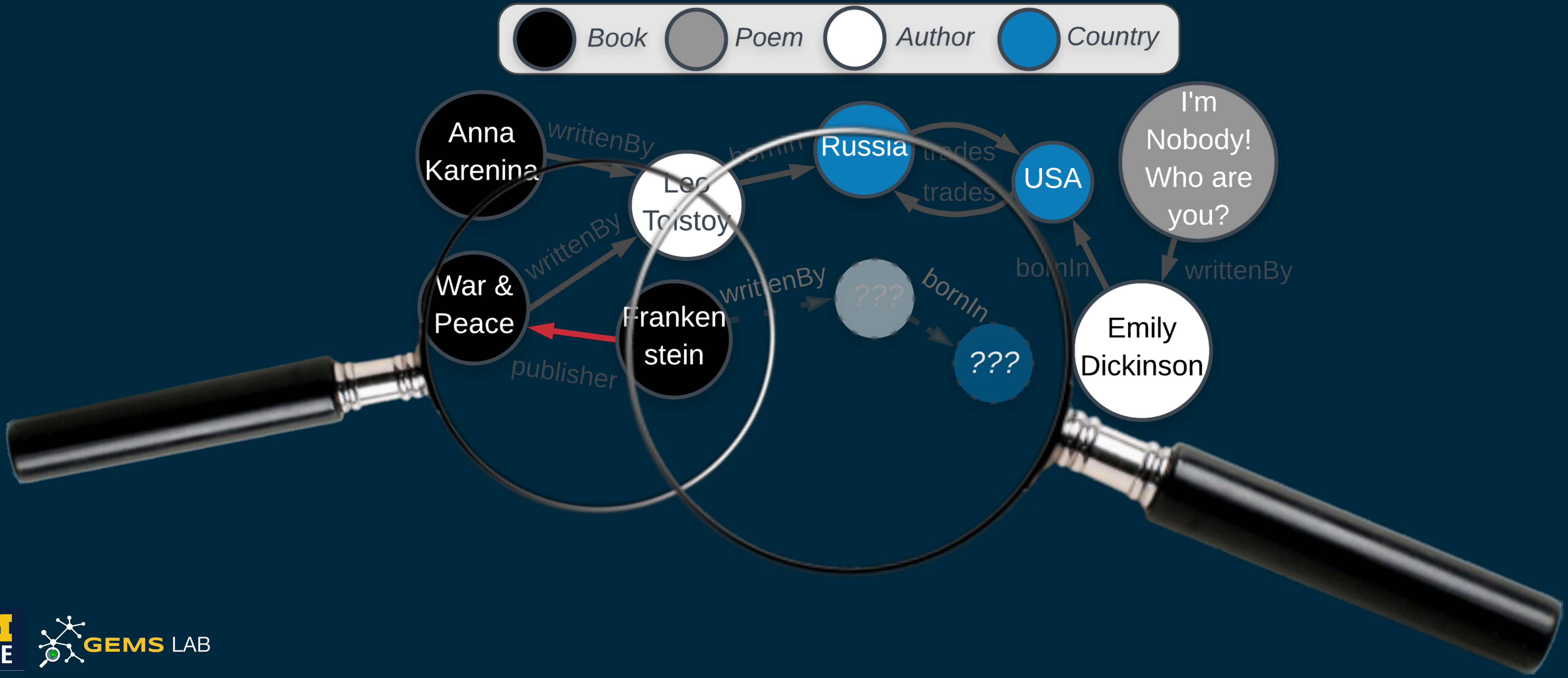
# This talk: Knowledge Graph Completion

- Evaluation of knowledge graph embeddings for trustworthy link prediction [EMNLP'20a]
- CoDEx: knowledge graph completion benchmark [EMNLP'20b]
- Knowledge graph summarization for unified error detection and completion [WWW'20]

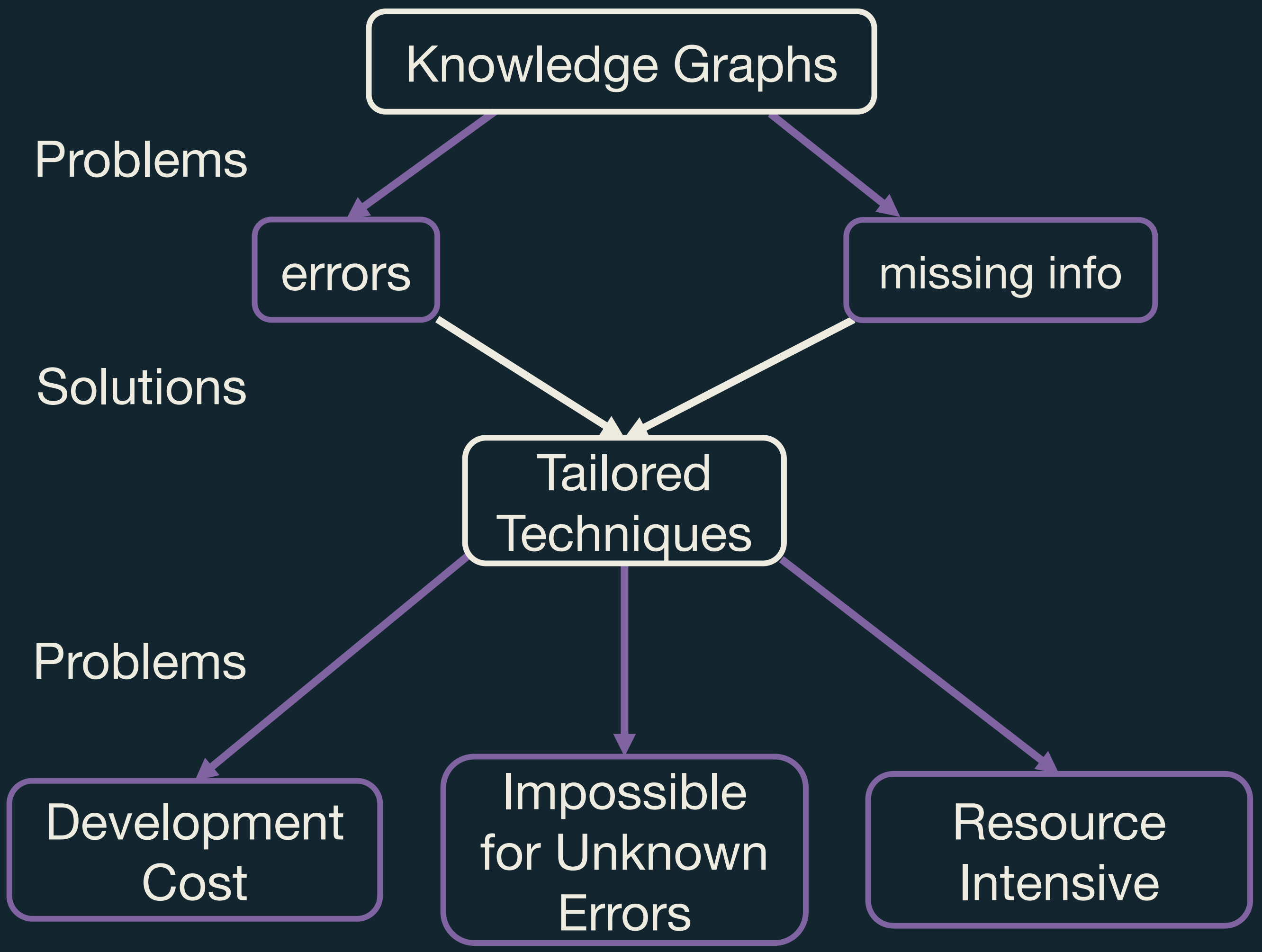



GEMS LAB

# Reminder:
# KGs have both errors & missing information



# Current Approach



Caleb Belth

Carol Zheng

Jilles Vreeken

## Current Approach

Knowledge Graphs

Problems

errors

missing info

Solutions

Tailored Techniques

Problems

Development Cost

Impossible for Unknown Errors

Resource Intensive

## Proposed Approach

Knowledge Graphs

Problem

abnormal: errors & missing info

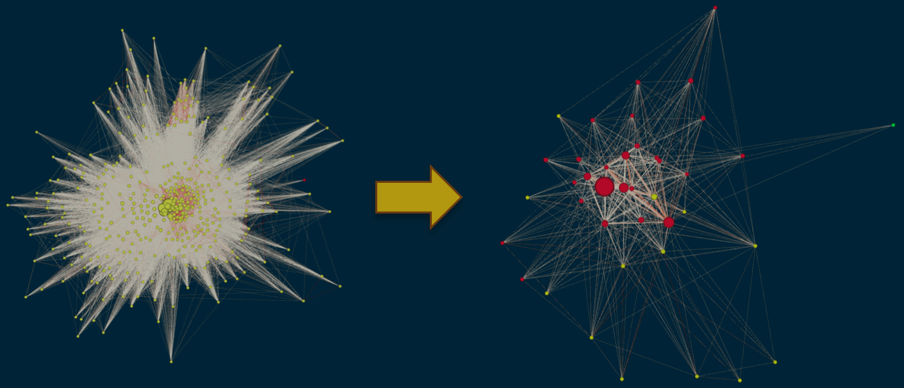
Solution

inductive summarization

# What is graph summarization?

Graph summarization seeks to find:

- a short representation of the input graph,
  - ✧ often in the form of an aggregated or sparsified graph, or a set of structures
- which reveals patterns in the original data and preserves specific structural or other properties, depending on the application domain.



## Graph Summarization Methods and Applications: A Survey

YIKE LIU, TARA SAFAVI, ABHILASH DIGHE, and DANAI KOUTRA, University of Michigan, Ann Arbor

While advances in computing resources have made processing enormous amounts of data possible, human ability to identify patterns in such data has not scaled accordingly. Efficient computational methods for condensing and simplifying data are thus becoming vital for extracting actionable insights. In particular, while data summarization techniques have been studied extensively, only recently has summarizing interconnected data, or *graphs*, become popular. This survey is a structured, comprehensive overview of the state-of-the-art methods for summarizing graph data. We first broach the motivation behind and the challenges of graph summarization. We then categorize summarization approaches by the type of graphs taken as input and further organize each category by core methodology. Finally, we discuss applications of summarization on real-world graphs and conclude by describing some open problems in the field.



CCS Concepts: • **Mathematics of computing** → **Graph algorithms**; • **Information systems** → **Data mining**; **Summarization**; • **Human-centered computing** → *Social network analysis*; • **Theory of computation** → *Unsupervised learning and clustering*; • **Computing methodologies** → *Network science*;

Additional Key Words and Phrases: Graph mining, graph summarization

**ACM Reference format:**
Yike Liu, Tara Safavi, Abhilash Dighe, and Danai Koutra. 2018. Graph Summarization Methods and Applications: A Survey. *ACM Comput. Surv.* 51, 3, Article 62 (June 2018), 34 pages.
https://doi.org/10.1145/3186727

### 1 INTRODUCTION

As technology advances, the amount of data that we generate and our ability to collect and archive such data both increase continuously. Daily activities like social media interaction, web browsing, product and service purchases, itineraries, and wellness sensors generate large amounts of data, the analysis of which can immediately impact our lives. This abundance of generated data and its velocity call for data summarization, one of the main data mining tasks.

Since summarization facilitates the identification of structure and meaning in data, the data mining community has taken a strong interest in the task. Methods for a variety of data types

Y. Liu and T. Safavi contributed equally to this article.

This material was based on work supported in part by the National Science Foundation under grant IIS 1743088, Trove, and the University of Michigan. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation or other funding parties. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation here on.

Authors' addresses: Y. Liu, T. Safavi, A. Dighe, and D. Koutra, Bob and Betty Beyster Building, 2260 Hayward St, Ann Arbor, MI 48109; emails: {yikeliu, tsafavi, adighe, dkoutra}@umich.edu.


https://doi.org/10.1145/3186727





[Yike Liu, Tara Safavi, Abhilash Dighe, Danai Koutra. ACM Computing Surveys '18]

# KGIST: Knowledge Graph Inductive SummarizaTion

Given: a KG *G*

Find: a concise summary of *G*, consisting of inductive, soft rules.

Rules:
Normal

Exceptions & Unexplained:
Abnormal

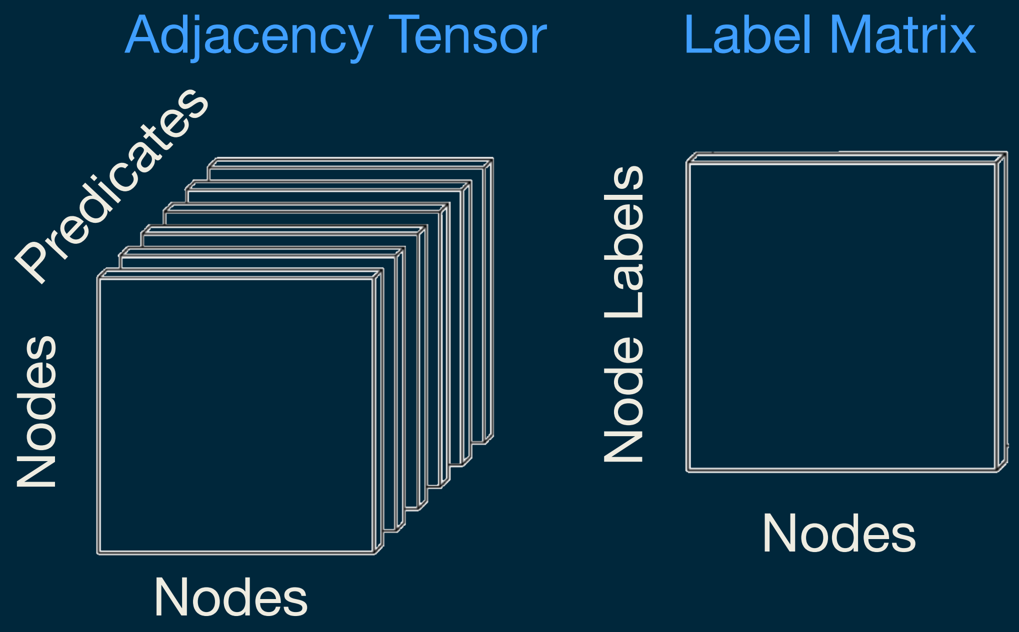
Key ideas:

1. Flipping the problem to unify refinement tasks
2. MDL-based approach for a concise set of rules

# Knowledge graph: Definition

Knowledge graph $G$ is a labeled, directed graph.

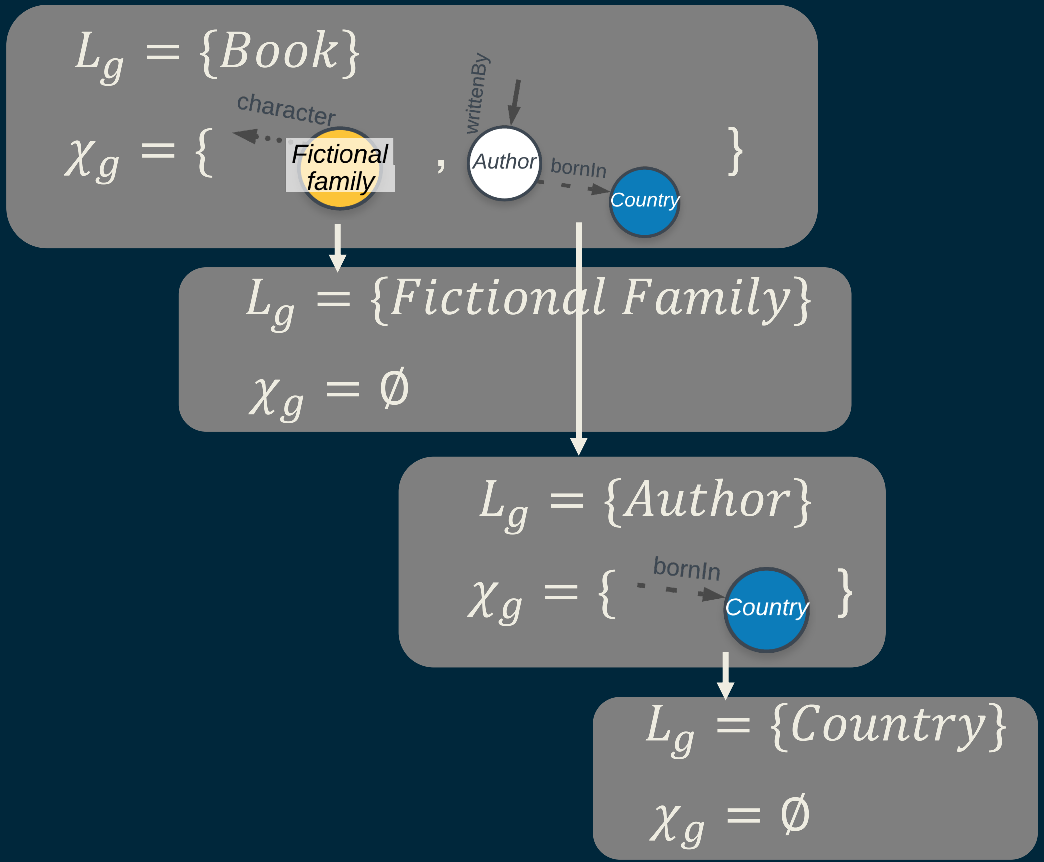
- edge = triple (subject node, predicate or relation, object node)
- Represented as:

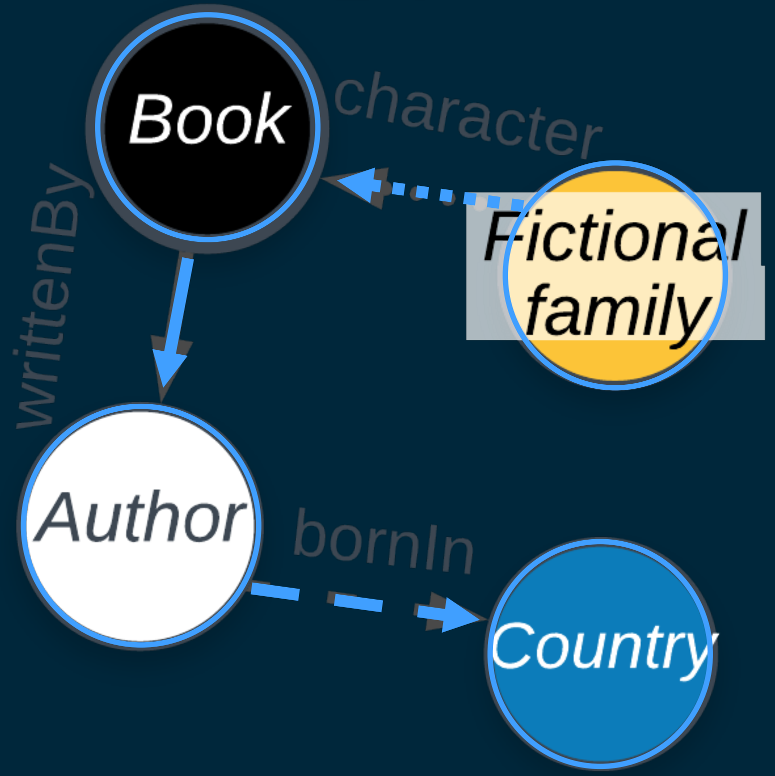



# Proposed Rule Definition: $g = (L_g,\ \chi_g)$

We formulate rules recursively as rooted, directed, and labeled graphs

- A rule asserts things about nodes with the root labels, $L_g$



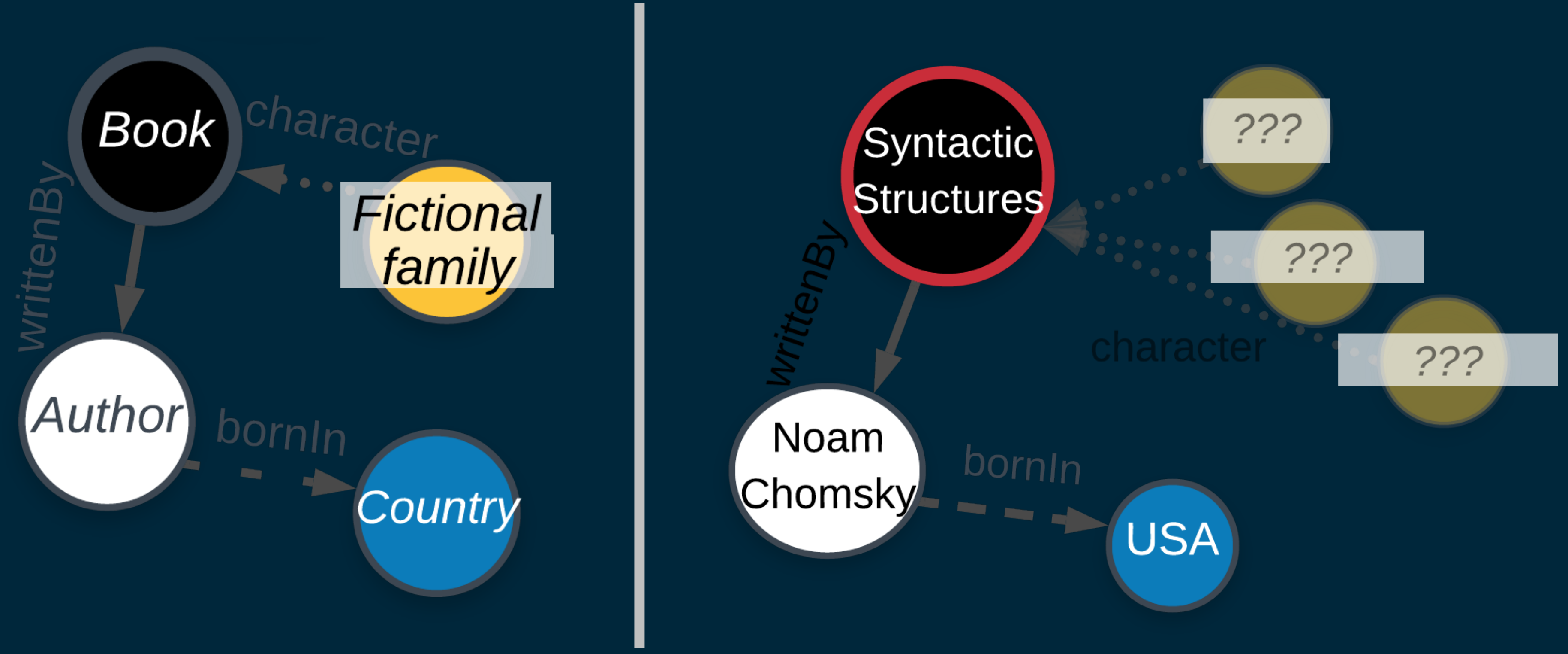



# The correct assertions, $\mathcal{A}_c^{(g)}$, of a rule

are guided traversals, which induce/instantiate subgraphs in the KG.

# The exceptions to a rule, $\mathcal{A}_{\xi}^{(g)}$

are failed guided traversals.



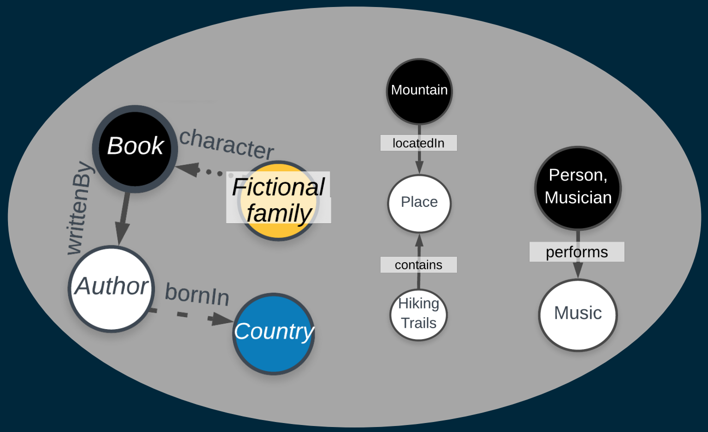
# KGIST: Knowledge Graph Inductive SummarizaTion

Given: a KG $G$

Find: a concise set of inductive rules $M$ that

$$\min L(G,M) = \underbrace{L(M)}_{\text{bits to describe } M} + \underbrace{L(G|M)}_{\text{bits to describe } G \text{ with } M}$$

$M =$




Model $M$: a set of rules

(each with correct assertions)

Normal

Expensive Parts of $L(G,M)$

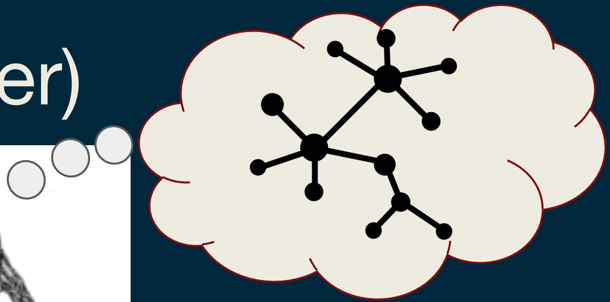
Abnormal

# Deriving $L(G, M)$: Idea

$L(G,M) = L(M) + L(G|M)$

- Take description length literally
- How many bits to describe a KG?

Alice (sender)



Hey Alice, could you tell me about your KG?

Bob (receiver)

# MDL Model: Overview

Alice

Sure! I'll send:

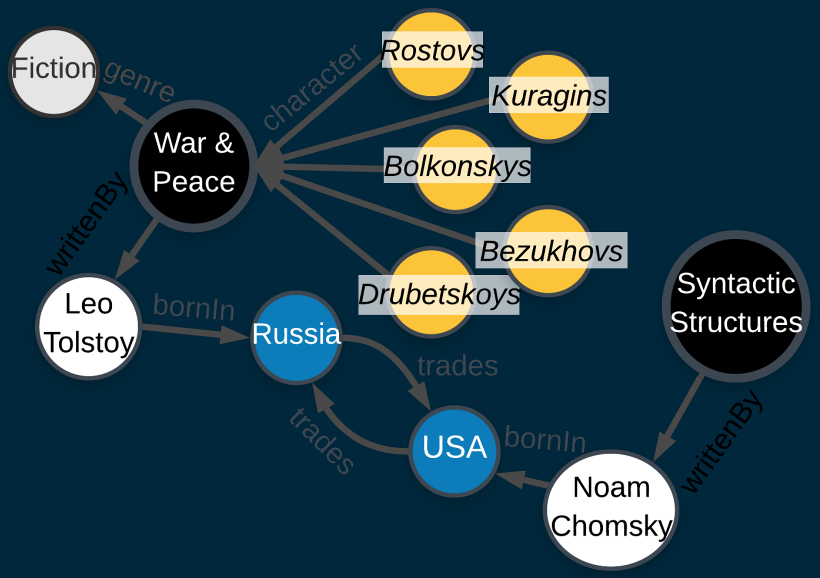
1) Model-independent information
2) A model
3) Any error the model makes

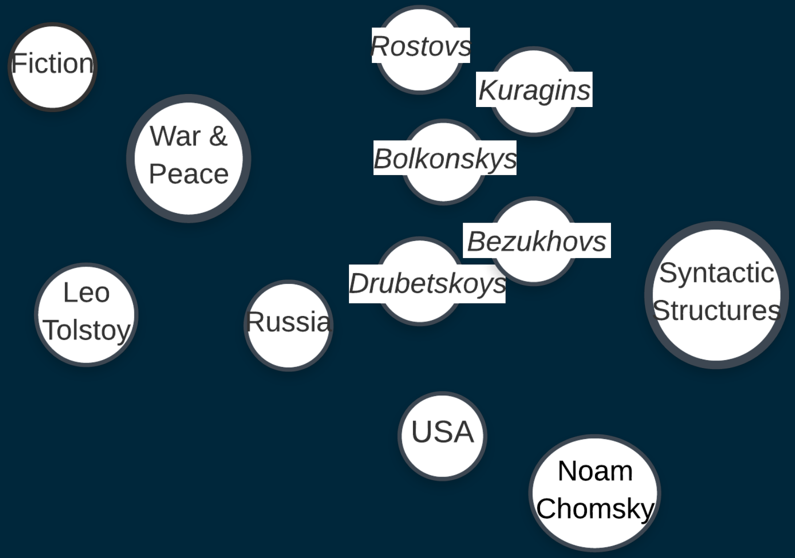
Bob

Ok, send the model the minimizes

$$L(G, M) = L(M) + L(G|M)$$

# MDL Model: $L(G, M) = L(M) + L(G|M)$



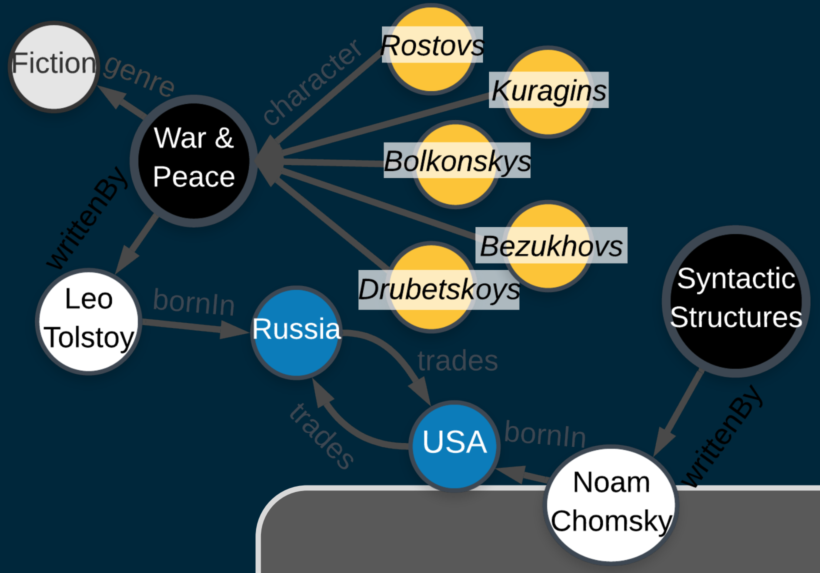



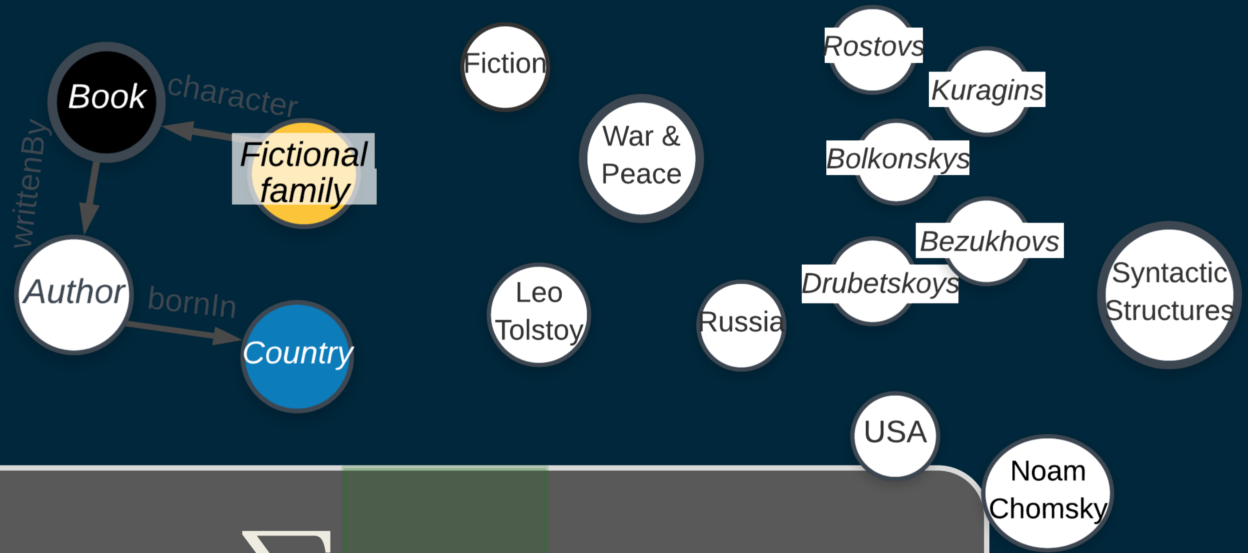

Alice

Model independent info:
# nodes, # edges, node ids …

Bob

# MDL Model: $L(G, M) = L(M) + L(G|M)$



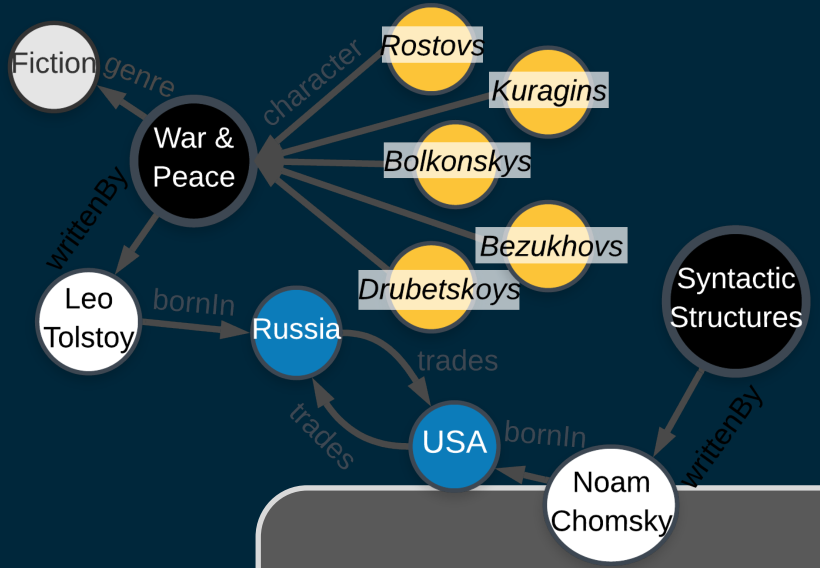



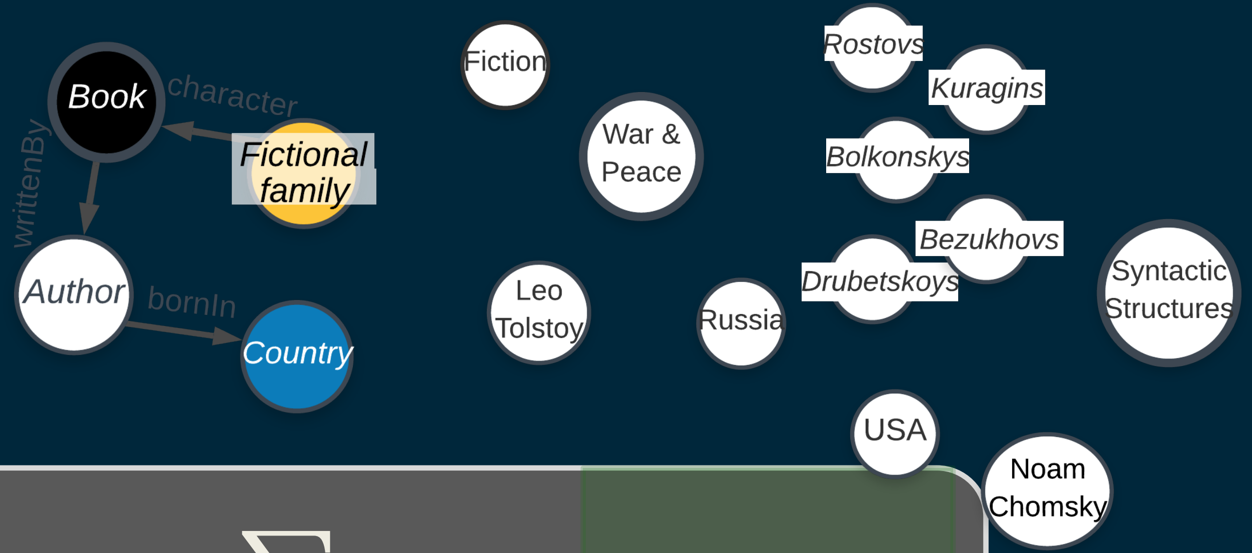

Alice

$$L(M) = \underbrace{\log(2 * |L_{\mathcal{V}}|^2 + |L_{\mathcal{E}}| + 1)}_{\#\text{ rules}} + \sum_{g \in M} (\underbrace{L(g)}_{\text{rule}} + \underbrace{L(\mathcal{A}^{(g)})}_{\text{assertions}})$$

Bob

# MDL Model: $L(G, M) = L(M) + L(G|M)$




Alice

$$L(M) = \underbrace{\log(2 * |L_{\mathcal{V}}|^2 + |L_{\mathcal{E}}| + 1)}_{\text{\# rules}} + \sum_{g \in M} \left(\underbrace{L(g)}_{\text{rule}} + \underbrace{L(\mathcal{A}^{(g)})}_{\text{assertions}}\right)$$

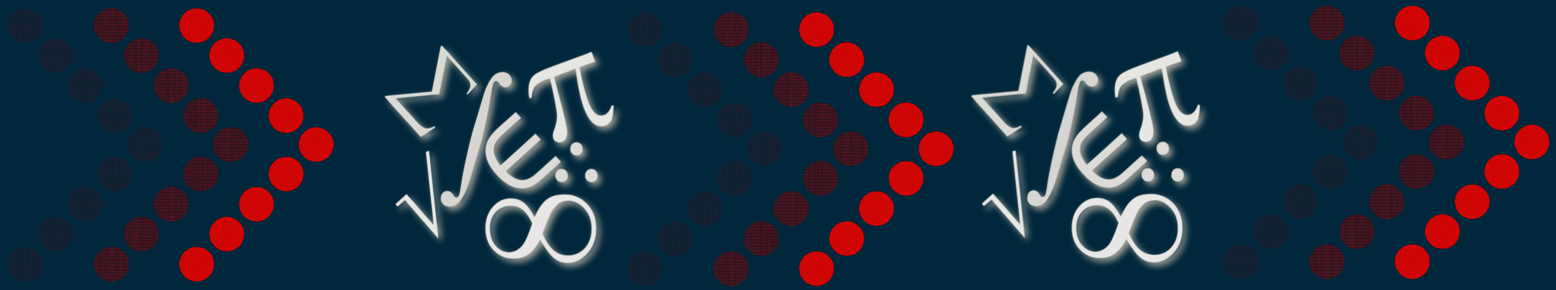
Bob

Great, now let me find all the books!

# MDL Model: $L(G, M) = L(M) + L(G|M)$

- Alice continues with the assertions, traversals etc…

Alice



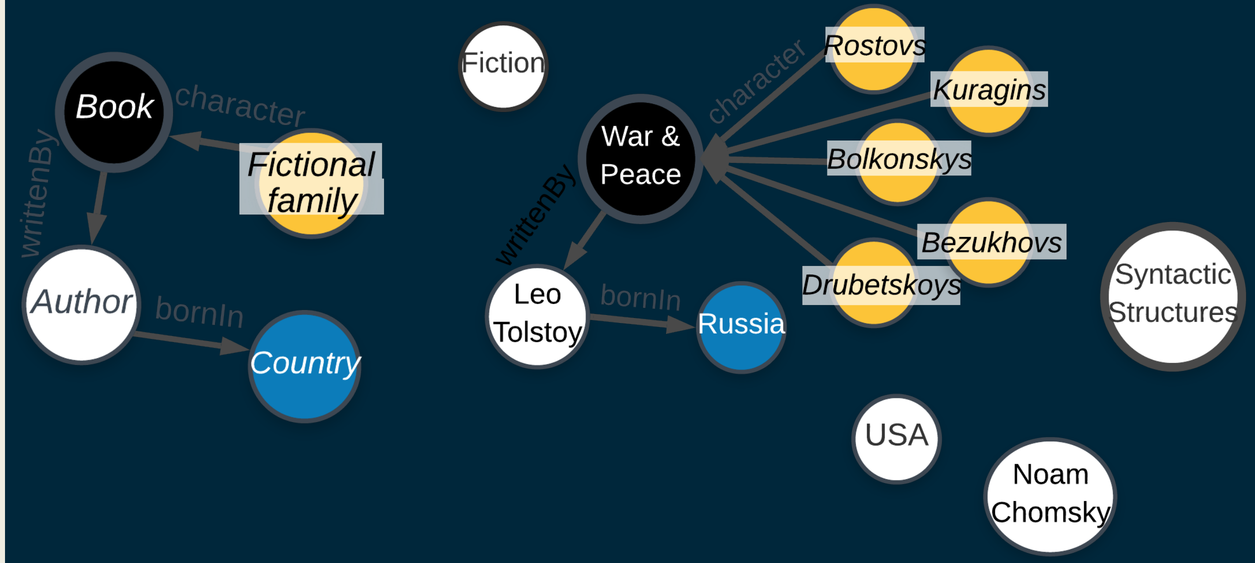
Bob

# MDL Model: $L(G, M) = L(M) + L(G|M)$

$$L(G|M) = \mathrm{L}(\boldsymbol{L}^{-}) + L(\boldsymbol{A}^{-})$$

I'll send the 1s in $\boldsymbol{L}$ and $\boldsymbol{A}$ that the rules didn't reveal

Alice



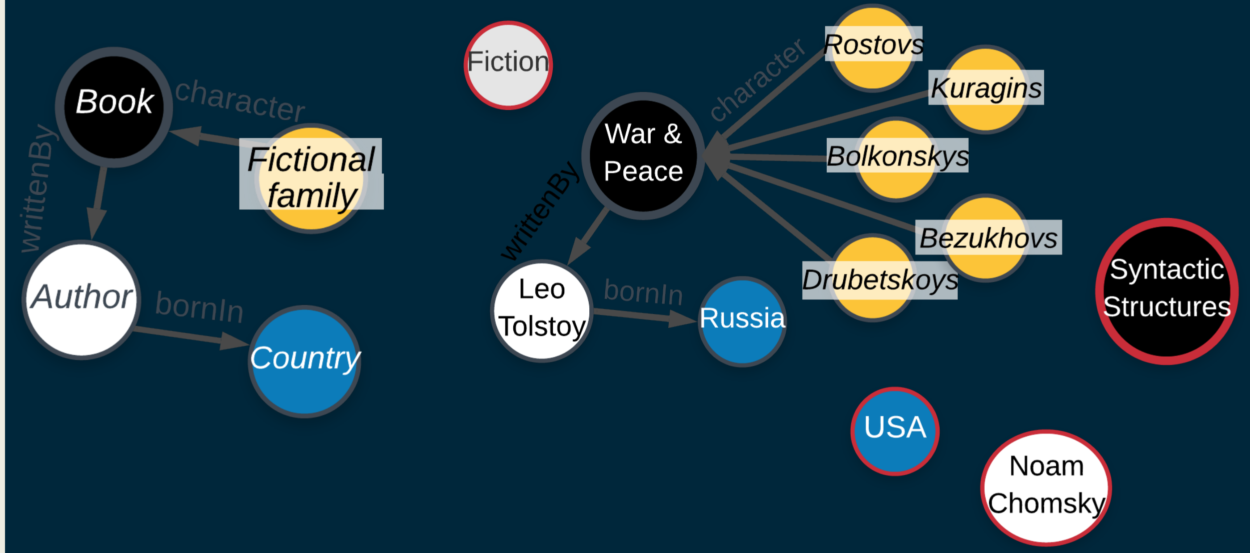

Bob

[Caleb Belth, Xinyi Zheng, et al. WWW '20]

github.com/GemsLab/KGIST

# MDL Model: $L(G, M) = L(M) + L(G|M)$

$$L(G|M) = \mathrm{L}(\boldsymbol{L}^{-}) + L(\boldsymbol{A}^{-})$$



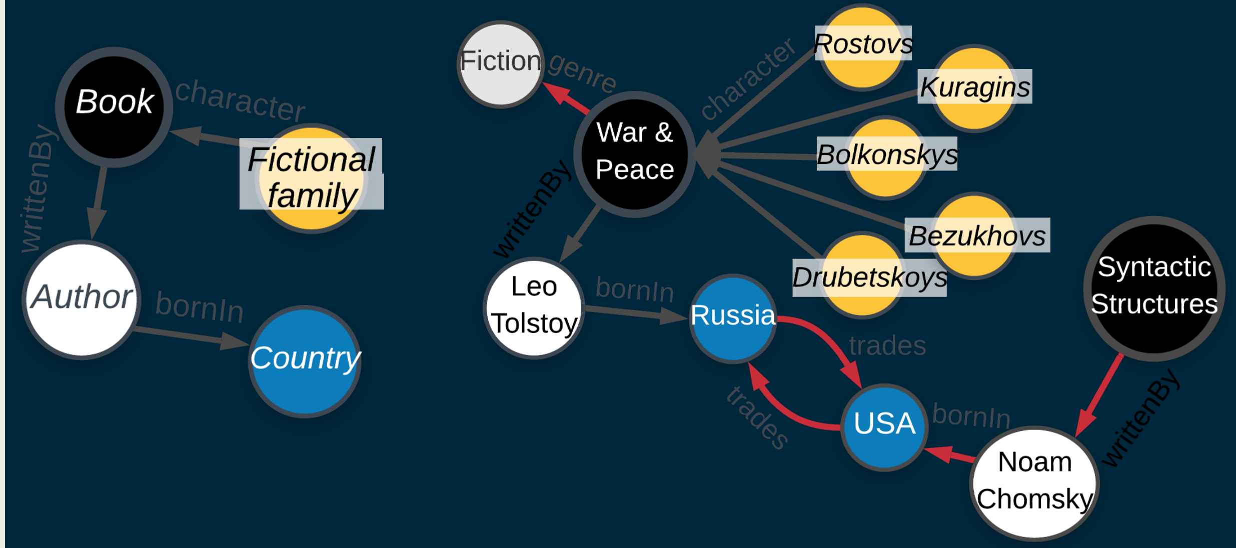

Alice

Bob

# MDL Model: $L(G, M) = L(M) + L(G|M)$

$$L(G|M) = \mathrm{L}(\boldsymbol{L}^{-}) + L(\boldsymbol{A}^{-})$$

Alice




Bob



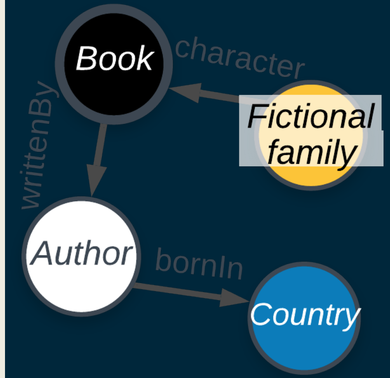
[Caleb Belth, Xinyi Zheng, et al. WWW '20]

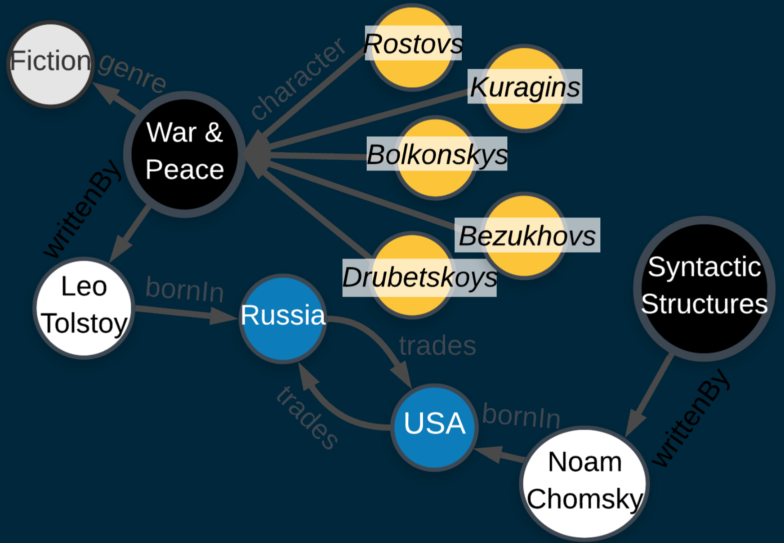
github.com/GemsLab/KGIST

# MDL Model: $L(G, M) = L(M) + L(G|M)$

There you go!

Alice

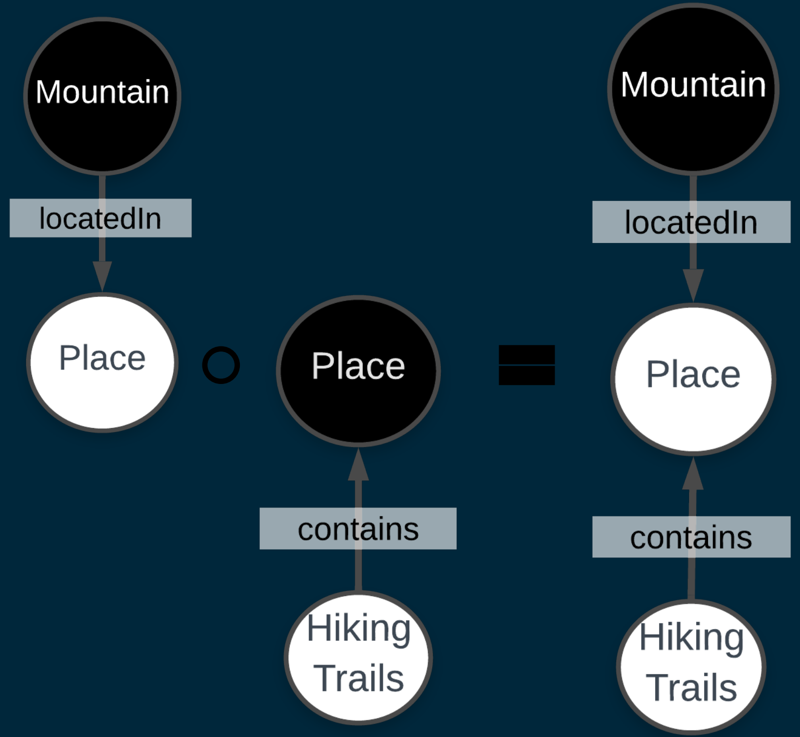



Bob

# KGIST Method: Overview

1. Generate candidate rules
2. Rank candidate rules
   - ✧ Based on how much they help explain/compress the KG
3. Select rules
   - ✧ Based on minimizing L(G,M)
4. Refine rules
   - ✧ Merging and nesting



# KGIST Anomaly Scores

- Anomalous entities: violate many rules
  - ✧ *MDL intuition*: many bits to describe a node as an exception

- Anomalous triples: unexplained edges, with anomalous endpoints

$$\eta(s,p,o) = \underbrace{\eta(s)+\eta(o)}_{\text{node endpoints}} + \underbrace{\eta^{(p)}(s,p,o)}_{\text{predicate}}$$

Alice

Bob

[Caleb Belth, Xinyi Zheng, et al. WWW '20]

github.com/GemsLab/KGIST

# Q1. Does KGIST find what is strange?

| | Supervised | | Unsupervised | | | |
|---|---|---|---|---|---|---|
| *Metric* | ComplEx | TransE | SDValidate | AMIE+ | KGIST_FREQ | KGIST+m |
| AUC | | | | | | |
| P@100 | | | | | | |
| R@100 | | | | | | |
| F1@100 | | | | | | |

Select q% of all nodes and,

$\lesssim 0.0188$

$\lesssim 0.0369$

remove label

billionaire, ~~entrepreneur~~, person

Bill Gates

add label

building, fruit

Taj Mahal

inject 1 or 2 edges

city

Des Moines

replace label

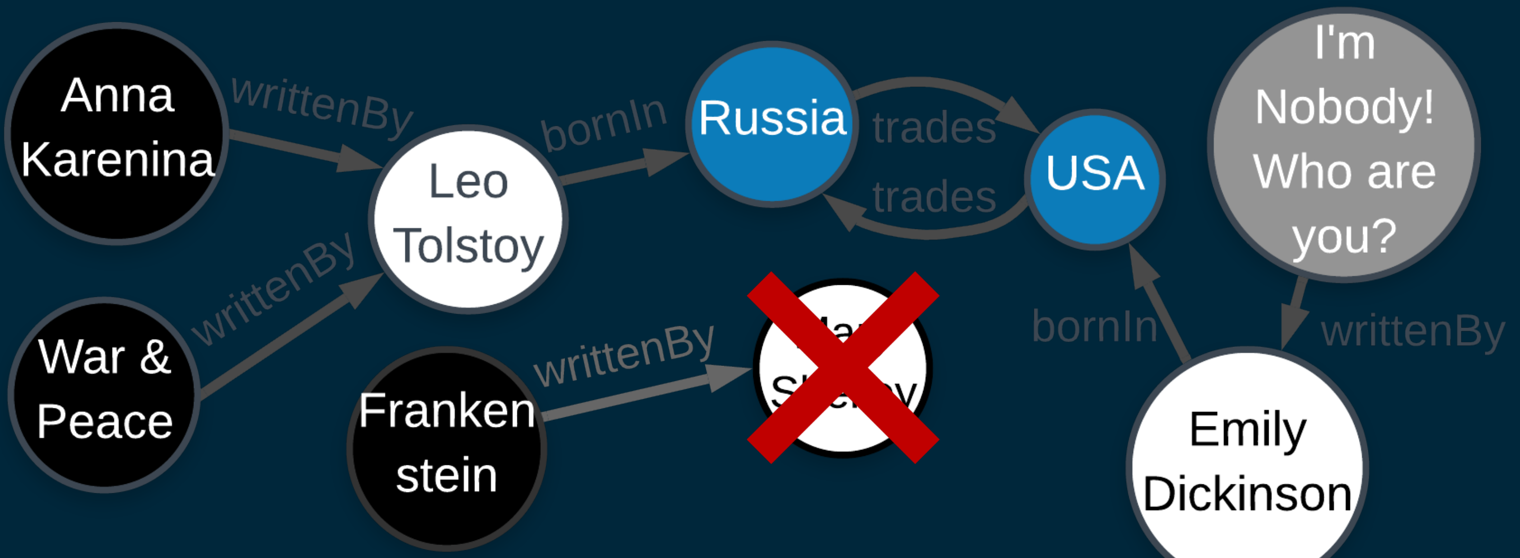
~~park,~~ car

Yellow Stone

KGIST performs best across all types of anomalies.

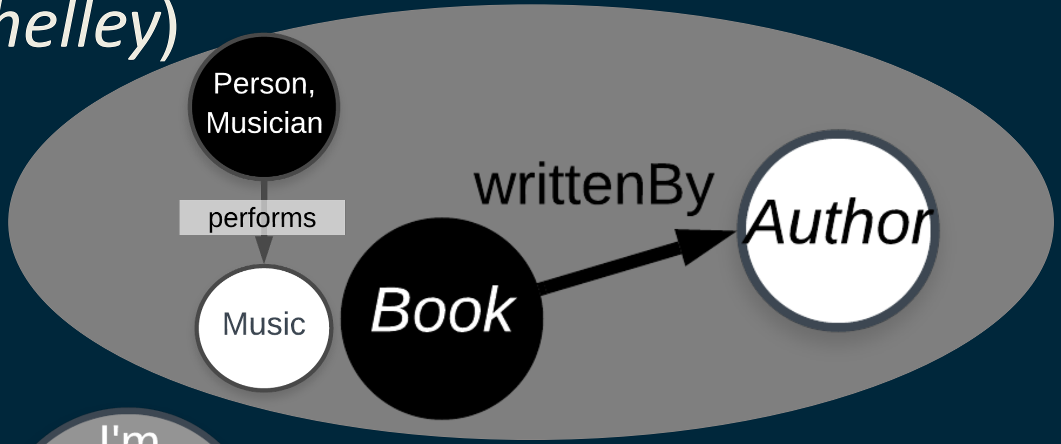
# Q2. Does KGIST find what is missing?

- Remove entities / nodes (e.g. *Mary Shelley*)



# Q2. Does KGIST find what is missing?

- Remove entities / nodes (e.g. *Mary Shelley*)
- Run KGIST on perturbed graph
- Find *where* entities are missing




# Q2. Does KGist find what is missing?

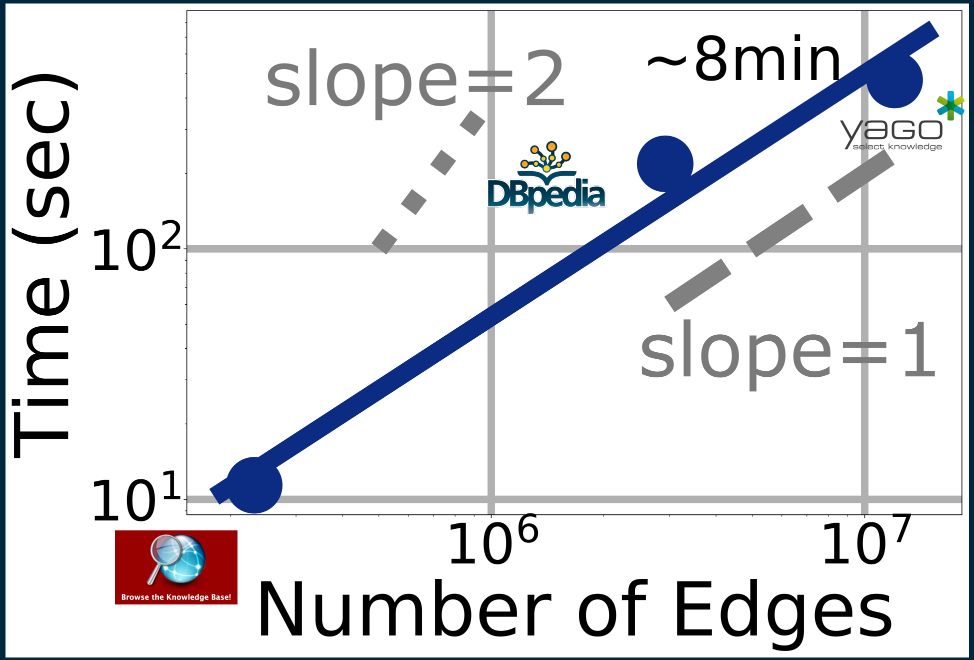
| | | Supervised | | Unsupervised | |
|---|---|---|---|---|---|
| *Dataset* | *Metric* | LP | AMIE+C [16] | Freq | KGist |
| NELL | R | N/A | 0.6587 ± 0.03 | 0.4589 ± 0.02 | 0.7598 ± 0.02 |
| | $R_L$ | N/A | N/A | 0.3924 ± 0.02 | 0.6636 ± 0.01 |
| DBpedia | R | N/A | 0.8187 ± 0.01 | 0.8049 ± 0.01 | 0.9288 ± 0.00 |
| | $R_L$ | N/A | N/A | 0.7839 ± 0.01 | 0.9179 ± 0.00 |


KGist significantly **outperforms** the baselines.
It **complements LP** methods.

[Caleb Belth, Xinyi Zheng, et al. WWW '20]

github.com/GemsLab/KGIST

# Q3. Is KGIST scalable?

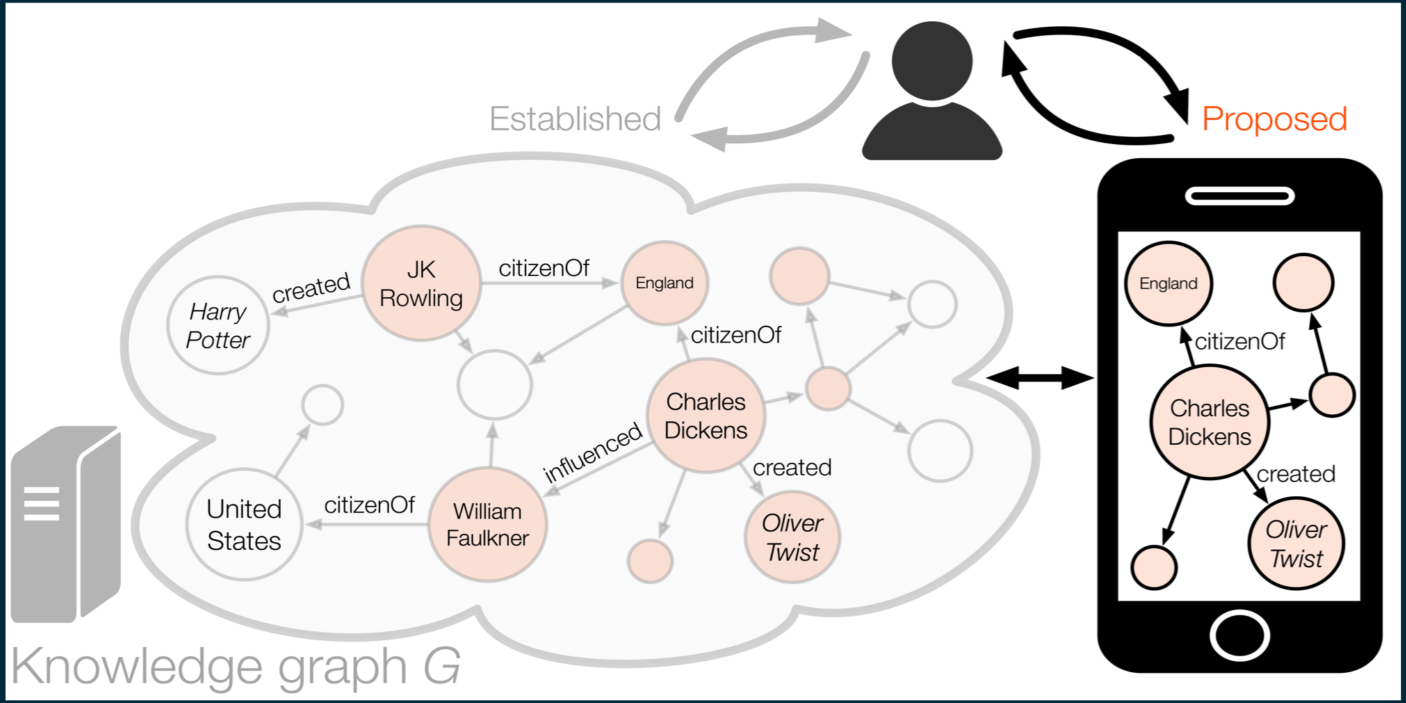



KGIST is near-linear
in the number of edges.

# Other types of summarization for KGs?

Personalized KG summarization for

private, offline, low-resource usage (e.g., QA)



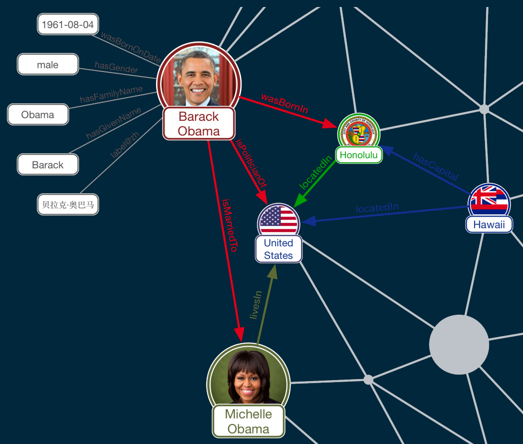

Personal summary of G



[Tara Safavi, Caleb Belth, et al. IEEE ICDM '19] https://github.com/GemsLab/GLIMPSE-personalized-KGsummarization

# Take-away messages: KG Completion

- Evaluation of trustworthiness of KGE-based link prediction through the lens of calibration [EMNLP'20a]
  - ✧ Standard models are overconfident in the open-world setting
  - ✧ Improving trustworthiness is harder than improving accuracy
- CoDEx: a new comprehensive dataset for knowledge graph completion [EMNLP'20a]
  - ✧ Improves upon existing benchmarks, fuses text and graph structure
  - ✧ Benchmarked on triple classification + link prediction: more discriminative power
- Rule-based summarization of KGs can help unify multiple refinement tasks that are traditionally solved by tailored approaches [WWW'20]
  - ✧ KG completion with KGIST: complementary to link prediction



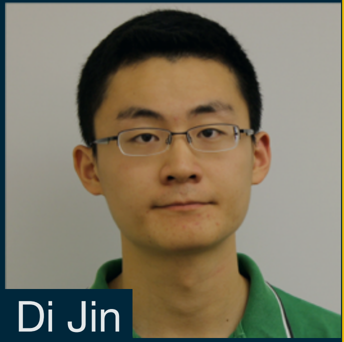
# Talk based on the following papers

- Tara Safavi, Danai Koutra, Edgar Meij. Evaluating the Calibration of Knowledge Graph Embeddings for Trustworthy Link Prediction. EMNLP 2020.
- Tara Safavi, Danai Koutra. CoDEx: A Comprehensive Knowledge Graph Completion Benchmark. EMNLP 2020.
- Caleb Belth, Xinyi (Carol) Zheng, Jilles Vreeken, Danai Koutra. What is normal, What is Strange, and What is Missing in a Knowledge Graph: Unified Characterization via Inductive Summarization. The Web Conference (WWW), 2020.
- Tara Safavi, Caleb Belth, Lukas Faber, Davide Mottin, Emmanuel Müller, Danai Koutra. Personalized Knowledge Graph Summarization: From the Cloud to Your Pocket. IEEE ICDM 2019.
- Y. Liu, T. Safavi, A. Dighe, D. Koutra. Graph Summarization Methods and Applications: A Survey. ACM Computing Surveys 2018.

## Thank you!
## Questions?

Danai Koutra
dkoutra@umich.edu
GEMS LAB



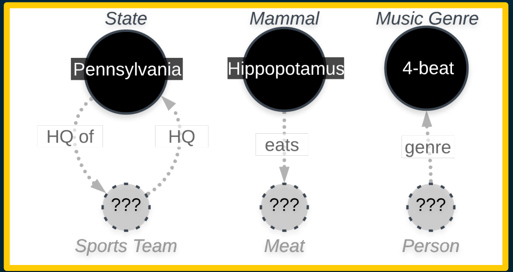






# Knowledge Graph Completion

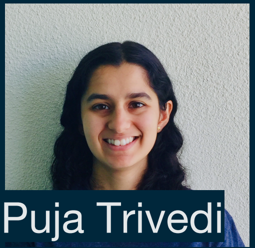
| Ranked triples predicted by KGE | Calib. scores | True? |
|---|---|---|
| 1. (Beyoncé, citizen, India) | ? | ✗ |
| 2. (Beyoncé, citizen, USA) | ? | ✓ |
| 3. (Beyoncé, citizen, jazz music) | ? | ✗ |
| ⋮ | ⋮ | ⋮ |

CoDEx




</>

github.com/tsafavi/codex
github.com/GemsLab/KGIST
github.com/GemsLab/GLIMPSE-personalized-KGsummarization



NSF ARL P&G MIDAS amazon Adobe Microsoft Azure