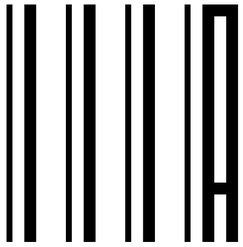
DIPARTIMENTO
DI INGEGNERIA
DELL'INFORMAZIONE



Efficient Evaluation of Knowledge Graph Quality

Challenges and Opportunities

Stefano Marchesin

09/05/2025

Tales on Data Science and Big Data

Outline

- Introduction
- Problem
- Efficient KG Evaluation
- Case Study
- Downstream Application
- Conclusions

Introduction

Empowering Access to Complex Knowledge Domains

A Knowledge Graph (KG):

- Captures both the structural and semantic aspects of information;
- Enables the representation, integration, and querying of complex knowledge domains.

Empowering Access to Complex Knowledge Domains

A Knowledge Graph (KG):

- Captures both the structural and semantic aspects of information;
- Enables the representation, integration, and querying of complex knowledge domains.

KGs store relational facts, in the form of (s, p, o) triples, where:

- (s)ubjects are entities or blank nodes;
- (o)bjects are entities, attributes, or blank nodes;
- (p)redicates express the relationships between them.

Knowledge Graphs: Preliminaries (1/2)

A Knowledge Graph (KG) is a directed, edge-labeled multi-graph $G = (V, R, \eta)$:

- $V = \{E \cup A \cup B\}$ is the set of nodes in G, where E are entities, A attributes and B blank nodes;
- R is the set of relationships between nodes in G;
- $\eta : R \rightarrow (E \cup B) \times (E \cup A \cup B)$ is a function assigning an ordered pair of nodes to each relationship.

Bonifati *et al.* Querying Graphs. 2018

Knowledge Graphs: Preliminaries (1/2)

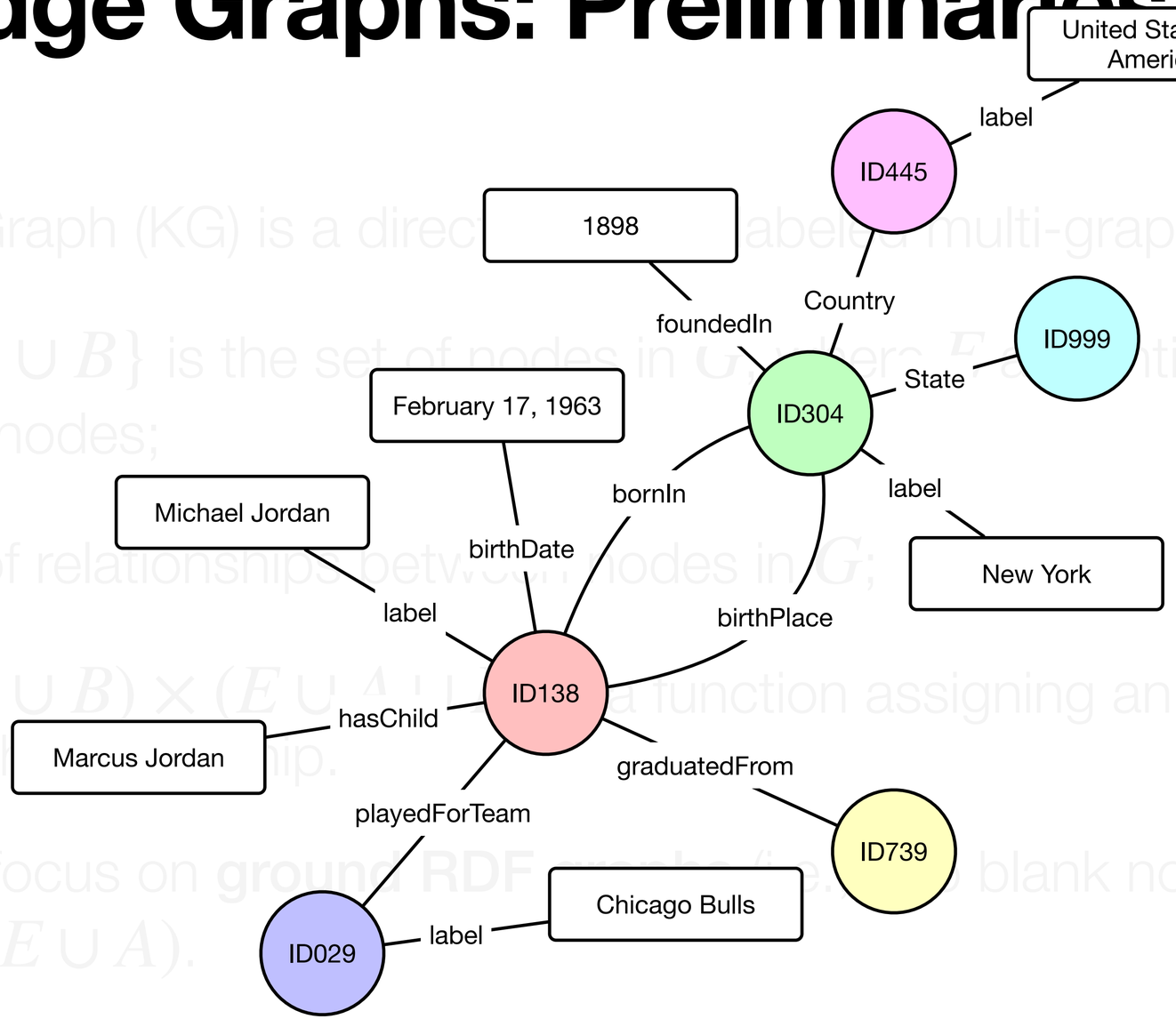
A Knowledge Graph (KG) is a directed, edge-labeled multi-graph $G = (V, R, \eta)$:

- $V = \{E \cup A \cup B\}$ is the set of nodes in G, where E are entities, A attributes and B blank nodes;
- R is the set of relationships between nodes in G;
- $\eta : R \rightarrow (E \cup B) \times (E \cup A \cup B)$ is a function assigning an ordered pair of nodes to each relationship.

In this talk, we focus on **ground RDF graphs** (i.e., w/o blank nodes), hence $\eta : R \rightarrow E \times (E \cup A)$.

Bonifati *et al.* Querying Graphs. 2018

Knowledge Graphs: Preliminaries (1/2)



A Knowledge Graph (KG) is a directed labeled multi-graph $G = (V, R, \eta)$:

- $V = \{E \cup A \cup B\}$ is the set of nodes in G, where E entities, A attributes and B blank nodes;
- R is the set of relationships between nodes in G;
- $\eta : R \rightarrow (E \cup B) \times (E \cup A \cup B)$ is a function assigning an ordered pair of nodes to each relationship.

In this talk, we focus on **ground RDF graphs** (i.e., no blank nodes), hence $\eta : R \rightarrow E \times (E \cup A)$.

Bonifati *et al.* Querying Graphs. 2018

Knowledge Graphs: Preliminaries (2/2)

The η function produces the ternary relation T of G.

Thus, the ternary relation T is the set of (s, p, o) triples such that $s \in E$, $p \in R$, and $o \in (E \cup A)$, where $M = |T|$ is its size.

Triples whose object is an entity are called **triples with entity property**.
Triples with attribute objects are known as **triples with data property**.

In this talk, we consider triples as first-class citizens next to nodes and relationships.Therefore, we define a KG as $G = (V, R, T, \eta)$.

Knowledge Graphs: Preliminaries (2/2)

The η function produces the ternary relation T of G.

Thus, the ternary relation T is the set of (s, p, o) triples such that $s \in E$, $p \in R$, and $o \in (E \cup A)$, where $M = |T|$ is its size.

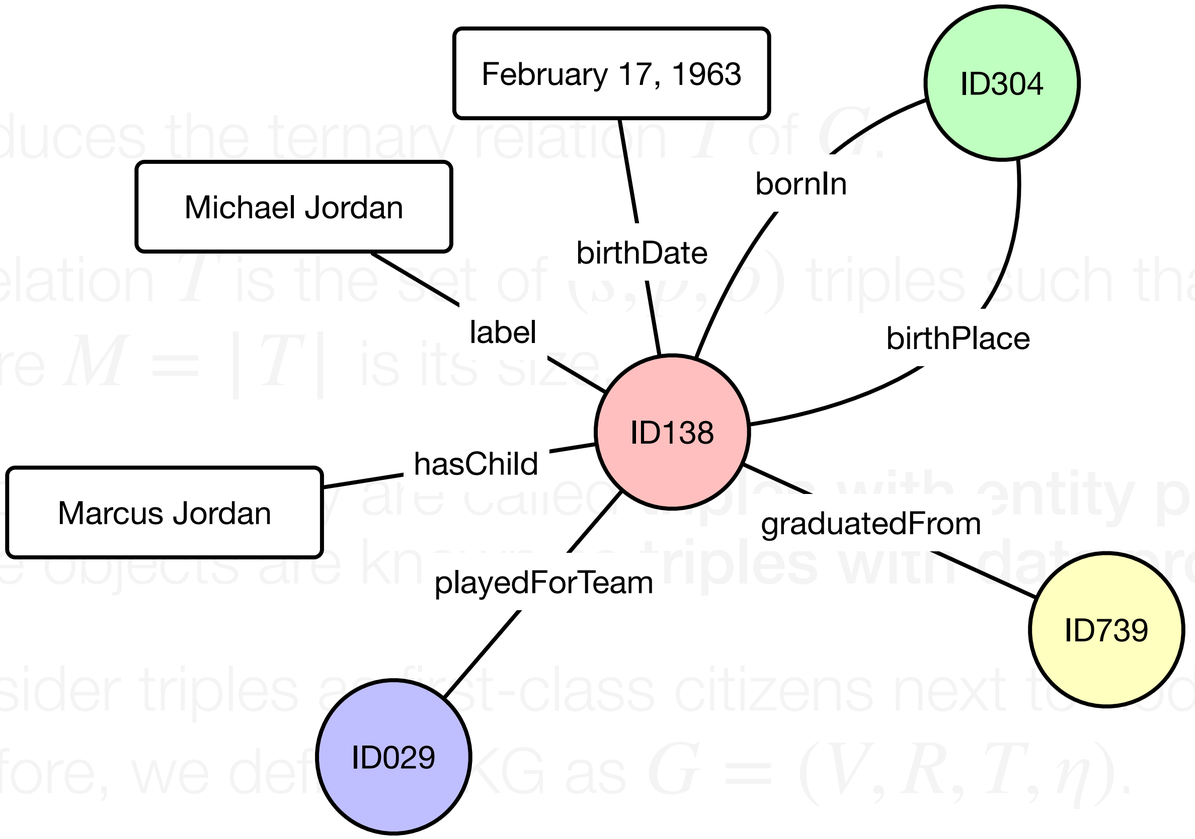
Triples whose object is an entity are called **triples with entity property**.
Triples with attribute objects are known as **triples with data property**.

In this talk, we consider triples as first-class citizens next to nodes and relationships.Therefore, we define a KG as $G = (V, R, T, \eta)$.

We also define an **entity cluster** $G[e] = \{(s, p, o) \in T \mid s = e\}$ as a set of triples in $T \in G$ sharing the same subject $e \in V$.

Bonifati *et al.* Querying Graphs. 2018

Knowledge Graphs: Preliminaries (2/2)



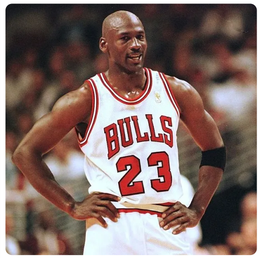
We also define an **entity cluster** $G[e] = \{(s, p, o) \in T \mid s = e\}$ as a set of triples in $T \in G$ sharing the same subject $e \in V$.

Knowledge Graphs: Downstream Tasks

KGs are pivotal resources for several downstream tasks: Information Retrieval, Item Recommendation, Question Answering

Knowledge Graphs: Downstream Tasks

KGs are pivotal resources for several downstream tasks: **Information Retrieval**, Item Recommendation, Question Answering

- KG data can be used to retrieve entity-centric information
  - e.g., Michael Jordan
- KG data can be used to create entity cards in SERPs
  - 

    About

    Michael Jeffrey Jordan, also known by his initials MJ, is an American businessman and former professional basketball player. His profile on the official National Basketball Association website states that "by acclamation, Michael Jordan is the greatest basketball player of all time." Wikipedia

    Born: February 17, 1963, Cumberland Hospital
    Net worth: 3 billion USD (2023) Forbes
    Spouse: Yvette Prieto (m. 2013), Juanita Vanoy (m. 1989–2006)
    Children: Marcus Jordan, Jeffrey Michael Jordan, Jasmine M. Jordan
    Number: 23 (Chicago Bulls / Shooting guard), MORE
    Teammates: Scottie Pippen, Magic Johnson, Larry Bird, MORE
    Parents: James R. Jordan, Sr., Deloris Jordan

Knowledge Graphs: The Need for Evaluation

Knowing the accuracy of KGs is fundamental for their **informed use** and to **activate correction mechanisms:**

- KG accuracy can be considered to favor search towards validated facts
- KG platforms are hampered by incorrect data: high-traffic queries on low-quality KG parts
- KG embeddings are sensitive to unreliable data
- KG data used to train LLMs: incorrect data leads to hallucinations, biases, and systematic errors

Ilyas *et al.* Saga: A Platform for Continuous Construction and Serving of Knowledge at Scale. SIGMOD. 2022

Pujara *et al.* Sparsity and Noise: Where Knowledge Graph Embeddings Fall Short. EMNLP. 2017

Reinanda *et al*. Knowledge Graphs: An Information Retrieval Perspective. Found. Trends Inf. Retr. 2020

Problem

It’s All about Scale

Objective: evaluate the accuracy of KGs to make an informed use of them.

Scenario: real-life KGs (e.g., YAGO, NELL, DBpedia, WikiData, etc.) are large-scale, from **hundreds of millions to billions** of triples.

Solution (naive): annotate every triple (or fact) stored within the KG.

It’s All about Scale

Objective: evaluate the accuracy of KGs to make an informed use of them.

Scenario: real-life KGs (e.g., YAGO, NELL, DBpedia, WikiData, etc.) are large-scale, from **hundreds of millions to billions** of triples.

Solution (naive): annotate every triple (or fact) stored within the KG.

Infeasible to **fully annotate** real-life KGs:

> Validating the entire DBpedia KG using Amazon Mechanical Turk would take more than **3,000 years**!

Paulheim and Bizer. Improving the Quality of Linked Data Using Statistical Distributions. International Journal on Semantic Web and Information Systems. 2014

Lack of Guarantees

Objective: evaluate the accuracy of KGs to make an informed use of them.

Scenario: real-life KGs (e.g., YAGO, NELL, DBpedia, WikiData, etc.) are large-scale, from **hundreds of millions to billions** of triples.

Solution (common practice): carry out manual annotations on a sample of the KG.

Lack of Guarantees

Objective: evaluate the accuracy of KGs to make an informed use of them.

Scenario: real-life KGs (e.g., YAGO, NELL, DBpedia, WikiData, etc.) are large-scale, from **hundreds of millions to billions** of triples.

Solution (common practice): carry out manual annotations on a sample of the KG.


Sample size is decided beforehand — is it large enough? biased? reproducible?


Confidence Intervals (**CIs**), required to quantify the uncertainties in the evaluation procedure, are not reported — how do we know if we are close to true value?

No guarantees of representativeness of the sample w.r.t. the entire KG.

Estimators to the Rescue

To **overcome the costs** required to fully evaluate KGs, and **provide statistical guarantees** to the evaluation procedure, **estimators** can be used.

Estimators $\hat{\mu}$ estimate $\mu(G)$ over a small sample, drawn via a sampling strategy $\mathcal{S}$.

To properly estimate the accuracy of G:

- Estimator $\hat{\mu}$ must be unbiased — that is, $E[\hat{\mu}] = \mu(G)$.
- Given the **point estimator** $\hat{\mu}$, a CI with a designated **confidence level** $1 - \alpha$ must be provided to quantify the uncertainties in the sampling procedure.

Problem Formulation: Constrained Minimization

Given:

- a KG G
- a sampling strategy $\mathcal{S}$
- an annotation cost function $\text{cost}(\cdot)$
- an upper bound ε for the Margin of Error (MoE) of the $1-\alpha$ CI

The problem of efficient KG accuracy evaluation can be defined as:

$$\underset{\mathcal{S}}{\text{minimize}}\ \text{cost}(\mathcal{S}(G))$$

$$\text{subject to}\ E[\hat{\mu}] = \mu(G),\ \text{MoE}(\hat{\mu}, \alpha) \leq \varepsilon$$

Gao *et al.* Efficient Knowledge Graph Accuracy Evaluation. VLDB. 2019

Problem Formulation: Constrained Minimization

Given:

- a KG G
- a sampling strategy $\mathcal{S}$
- an annotation cost function $\text{cost}(\cdot)$
- an upper bound ε for the Margin of Error (MoE) of the $1-\alpha$ CI

The problem of efficient KG accuracy evaluation can be defined as:

$$\underset{\mathcal{S}}{\text{minimize}}\ \text{cost}(\mathcal{S}(G))$$

$$\text{subject to}\ E[\hat{\mu}] = \mu(G),\ \text{MoE}(\hat{\mu}, \alpha) \leq \varepsilon$$

The sampling strategy draws the pool of triples that annotators audit for correctness.

Gao *et al.* Efficient Knowledge Graph Accuracy Evaluation. VLDB. 2019

Problem Formulation: Constrained Minimization

Given:

- a KG G
- a sampling strategy $\mathcal{S}$
- an annotation cost function $\text{cost}(\cdot)$
- an upper bound ε for the Margin of Error (MoE) of the $1-\alpha$ CI

The problem of efficient KG accuracy evaluation can be defined as:

$$\underset{\mathcal{S}}{\text{minimize}}\ \text{cost}(\mathcal{S}(G))$$

$$\text{subject to}\ E[\hat{\mu}] = \mu(G),\ \text{MoE}(\hat{\mu}, \alpha) \leq \varepsilon$$

The annotation cost represents the (total) time required to annotators for auditing the sample. Computational efficiency is not a direct concern, as long as the annotation time dominates.

Gao *et al.* Efficient Knowledge Graph Accuracy Evaluation. VLDB. 2019

Problem Formulation: Constrained Minimization

Given:

- a KG G
- a sampling strategy $\mathcal{S}$
- an annotation cost function $\text{cost}(\cdot)$
- an upper bound ε for the Margin of Error (MoE) of the $1-\alpha$ CI

Both the upper bound ε and the significance level α are user-defined parameters. The MoE represents half the width of a CI.

The problem of efficient KG accuracy evaluation can be defined as:

$$\underset{\mathcal{S}}{\text{minimize}}\ \text{cost}(\mathcal{S}(G))$$

$$\text{subject to}\ E[\hat{\mu}] = \mu(G),\ \text{MoE}(\hat{\mu}, \alpha) \leq \varepsilon$$

Gao *et al.* Efficient Knowledge Graph Accuracy Evaluation. VLDB. 2019

Problem Formulation: Constrained Minimization

Given:

- a KG G
- a sampling strategy $\mathcal{S}$
- an annotation cost function $\text{cost}(\cdot)$
- an upper bound ε for the Margin of Error (Mo

The problem of efficient KG accuracy evaluation

$$\underset{\mathcal{S}}{\text{minimize}}\ \text{cost}(\mathcal{S}(G))$$

$$\text{subject to}\ E[\hat{\mu}] = \mu(G),\ \text{MoE}(\hat{\mu}, \alpha) \leq \varepsilon$$

- Sampling strategy $\mathcal{S}$ should minimize $\text{cost}(\cdot)$
- MoE represents the stopping condition for convergence
- MoE depends on $\hat{\mu}$, which is based on the sampling $\mathcal{S}$
- $\hat{\mu}$ is unknown a priori

Gao *et al.* Efficient Knowledge Graph Accuracy Evaluation. VLDB. 2019

Efficient KG Evaluation

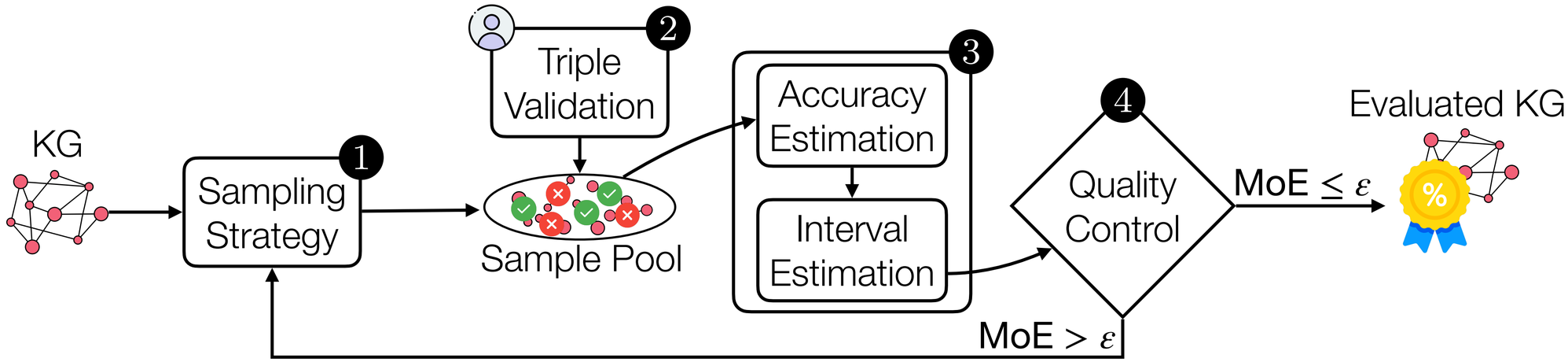
[1] Marchesin and Silvello. Efficient and Reliable Estimation of Knowledge Graph Accuracy. VLDB 2024

Optimization: Iterative Procedure

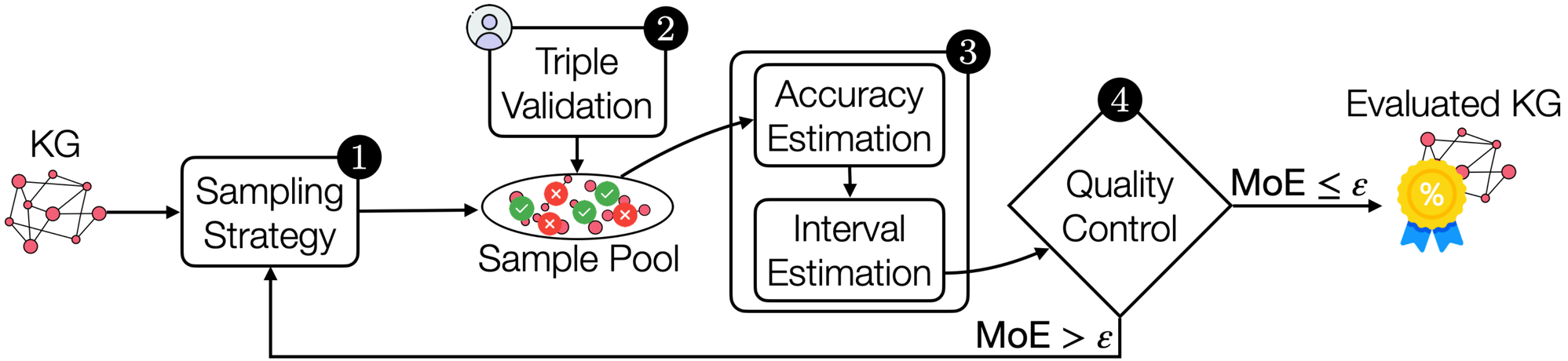
Since the problem depends on $\hat{\mu}$ to converge, an iterative procedure is well-suited to address it.

Such procedure samples and estimates iteratively and stops as soon as the MoE satisfies the specified threshold ε.

The approach prevents oversampling and unnecessary manual annotations, providing accurate estimations while minimizing costs.



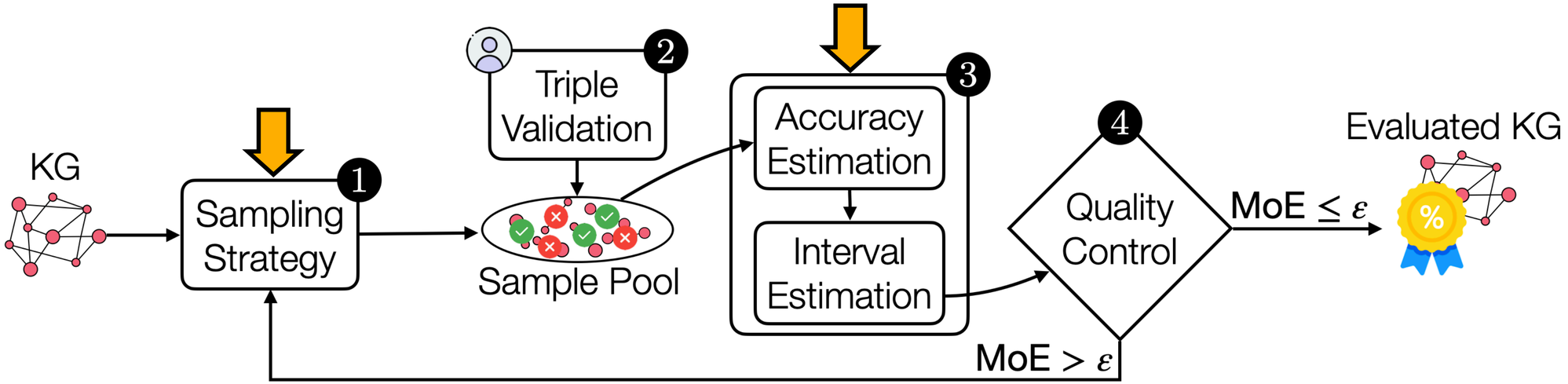
Research Focus: Interval Estimation



$$\underset{\mathcal{S}}{\text{minimize}}\ E[\text{cost}(\mathcal{S}(G))]$$

$$\text{subject to}\ E[\hat{\mu}] = \mu(G),\ \text{MoE}(\hat{\mu}, \alpha) \leq \varepsilon$$

Research Focus: Interval Estimation

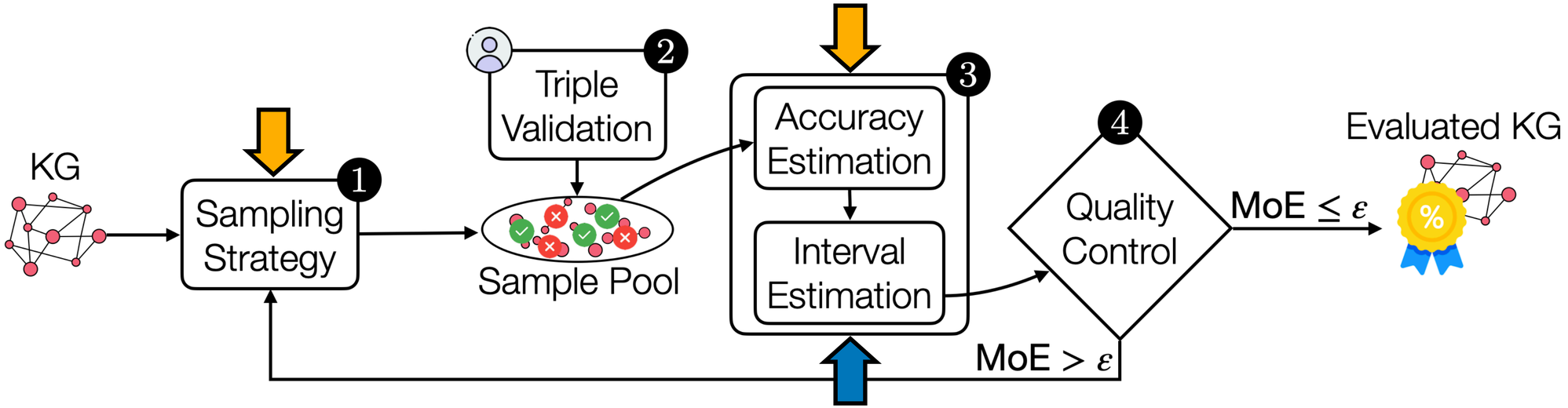


$$\underset{\mathcal{S}}{\text{minimize}}\ E[\text{cost}(\mathcal{S}(G))]$$

$$\text{subject to}\ E[\hat{\mu}] = \mu(G),\ \text{MoE}(\hat{\mu}, \alpha) \leq \varepsilon$$

Previous focus

Research Focus: Interval Estimation



$$\underset{\mathcal{S}}{\text{minimize}}\ E[\text{cost}(\mathcal{S}(G))]$$

$$\text{subject to}\ E[\hat{\mu}] = \mu(G),\ \text{MoE}(\hat{\mu}, \alpha) \leq \varepsilon$$

Previous focus

Our focus

Sampling and Accuracy Estimation: SOTA

Simple Random Sampling (SRS):

Draws n_T triples from G with uniform probability $1/M$

Accuracy Estimator:

$$\hat{\mu}_{SRS} = \frac{1}{n_T} \sum_{i=1}^{n_T} \mathbf{1}(t_i)$$

Sampling and Accuracy Estimation: SOTA

Simple Random Sampling (SRS):

Draws n_T triples from G with uniform probability $1/M$

Accuracy Estimator:

$$\hat{\mu}_{SRS} = \frac{1}{n_T} \sum_{i=1}^{n_T} \mathbf{1}(t_i)$$

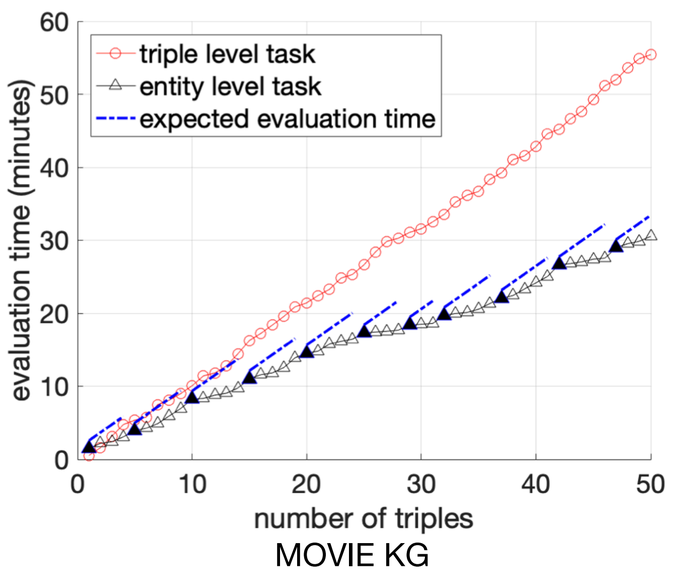
Two-stage Weighted Cluster Sampling (TWCS):

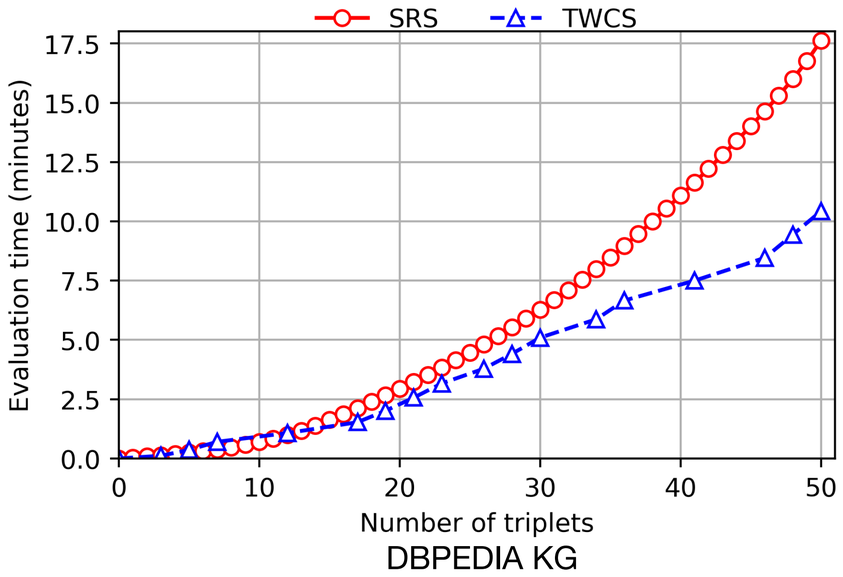
1) Draws n entity clusters with probability $\pi_i = \frac{M_i}{M}$ proportional to their sizes
2) Draws $\min\{M_i, m\}$ triples with SRS from each ith sampled cluster

$$\hat{\mu}_{TWCS} = \frac{1}{n} \sum_{i=1}^{n} \hat{\mu}_i$$

SRS VS TWCS: Annotation Cost Comparison

Annotating a new fact for an entity that has been already identified reduces the annotation cost compared to assess a new fact from unseen entities.





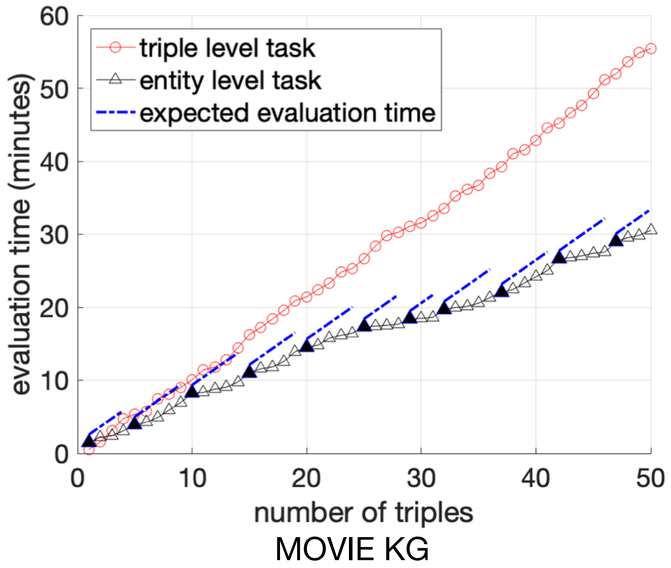
Gao *et al.* Efficient Knowledge Graph Accuracy Evaluation. VLDB 2019

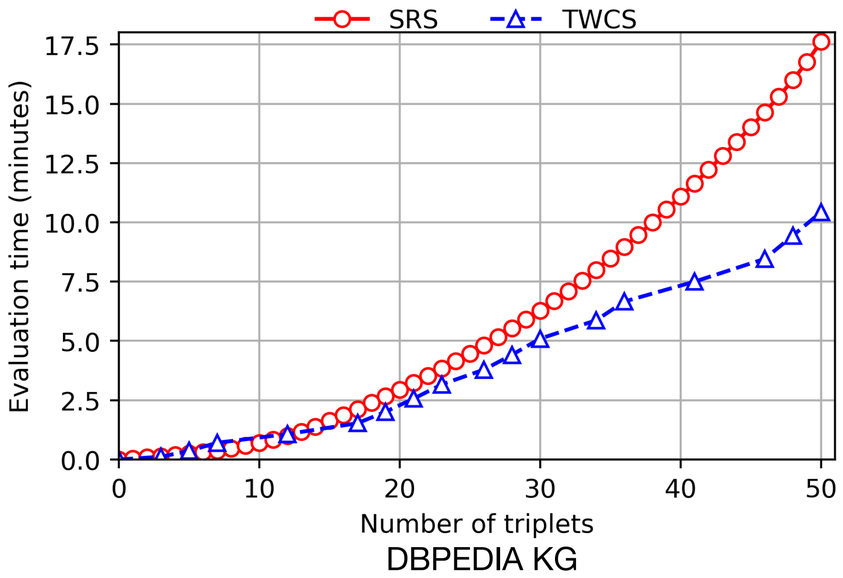
Marchesin *et al.* Utility-Oriented Knowledge Graph Accuracy Estimation with Limited Annotations: A Case Study on DBpedia. AAAI HCOMP 2024

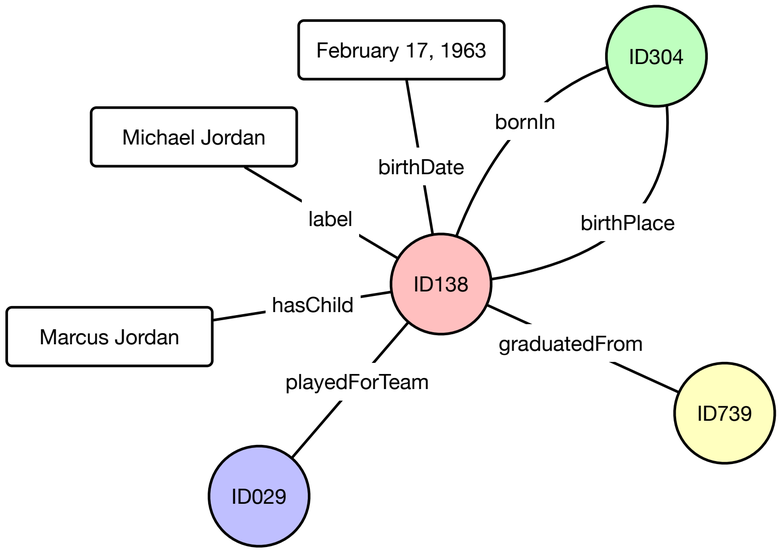
SRS VS TWCS: Annotation Cost Comparison

Annotating a new fact for an entity that has been already identified reduces the annotation cost compared to assess a new fact from unseen entities.

Entity clusters $G[e]$ become central aspects of the evaluation model.







Gao *et al.* Efficient Knowledge Graph Accuracy Evaluation. VLDB 2019

Marchesin *et al.* Utility-Oriented Knowledge Graph Accuracy Estimation with Limited Annotations: A Case Study on DBpedia. AAAI HCOMP 2024

Sampling and Accuracy Estimation: SOTA

Simple Random Sampling (SRS):

Draws n_T triples from G with uniform probability $1/M$

Accuracy Estimator:

$$\hat{\mu}_{SRS} = \frac{1}{n_T} \sum_{i=1}^{n_T} \mathbf{1}(t_i)$$

Two-stage Weighted Cluster Sampling (TWCS):

1) Draws n entity clusters with probability $\pi_i = \frac{M_i}{M}$ proportional to their sizes
2) Draws $\min\{M_i, m\}$ triples with SRS from each ith sampled cluster

$$\hat{\mu}_{TWCS} = \frac{1}{n} \sum_{i=1}^{n} \hat{\mu}_i$$

Stratified TWCS (STWCS):

1) Stratifies entity clusters into Q non overlapping strata
2) Applies TWCS within each qth stratum

$$\hat{\mu}_{STWCS} = \sum_{q=1}^{Q} W_q \cdot \hat{\mu}_{q,TWCS}$$

Confidence Interval: Wald Interval

Previous work relied on the Wald interval, which is obtained by **inverting the acceptance region** of the Wald large-sample normal test:

$$\left| \frac{(\hat{\mu} - \mu)}{\sqrt{V(\hat{\mu})}} \right| \leq z_{\alpha/2},$$

where $z_{\alpha/2}$ is the **critical value** of the standard normal distribution for a given significance level α.

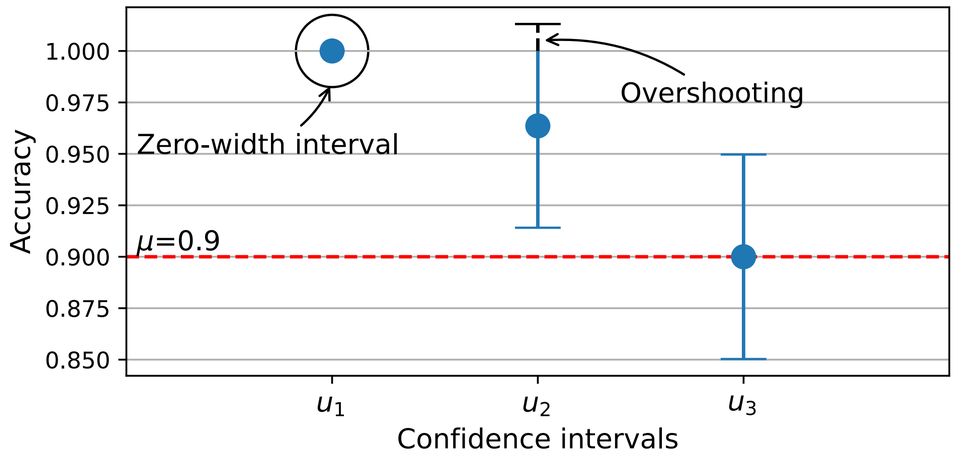
Leading to the well-known formula: $\hat{\mu} \pm z_{\alpha/2} \cdot SE(\hat{\mu}) \equiv \sqrt{V(\hat{\mu})}$

Limitations of the Wald interval

Serious issues when used on **(binomial) proportions**

Zero-width:
Implies absolute certainty
Underestimation of the error

Overshooting:
Exceeds [0, 1] boundaries
Approximation fails

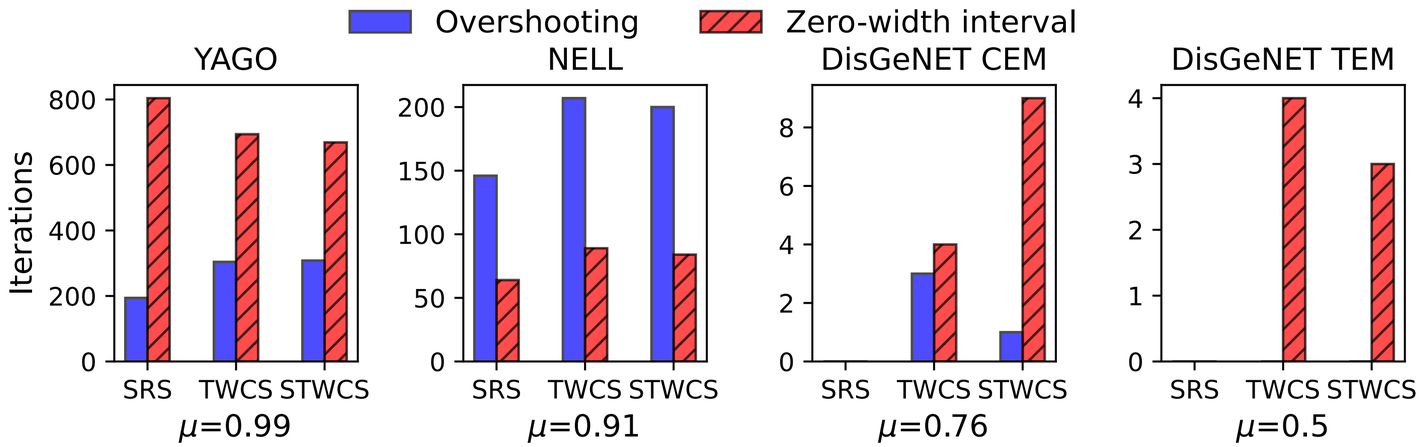


Limitations of the Wald interval

Serious issues when used on **(binomial) proportions**

Zero-width:
Implies absolute certainty
Underestimation of the error

Overshooting:
Exceeds [0, 1] boundaries
Approximation fails

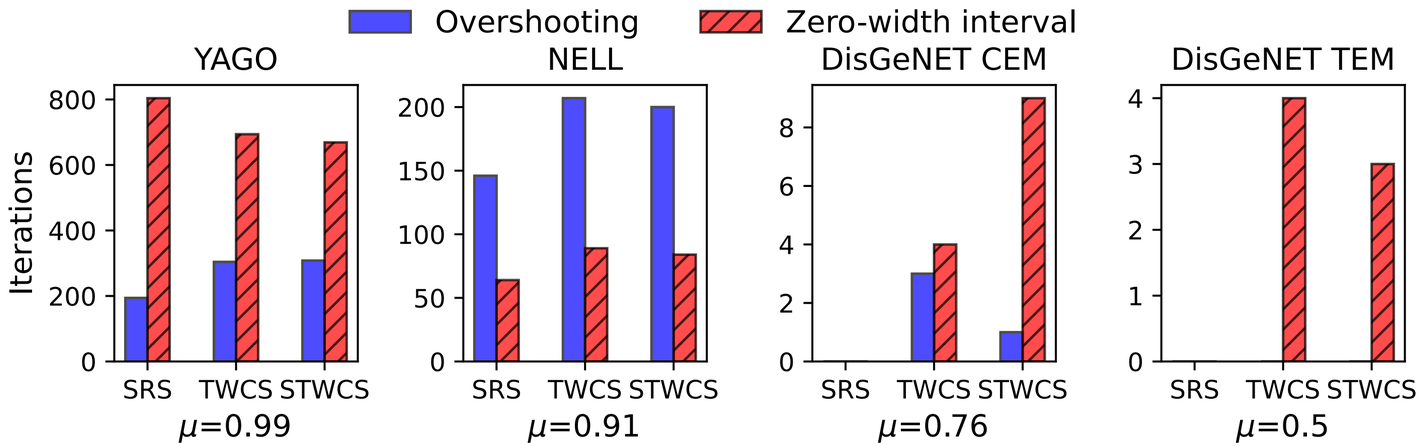


Limitations of the Wald interval

Serious issues when used on **(binomial) proportions**

Zero-width:
Implies absolute certainty
Underestimation of the error

Overshooting:
Exceeds [0, 1] boundaries
Approximation fails



Erratic coverage:

Fails to contain μ a number of $1 - \alpha$ times over repeated iterations (more on this later)
Reduces reliability

Binomial CIs: Overcoming Wald Limitations

To overcome/reduce the limitations of the Wald interval, Binomial CIs can be adopted.

These CIs are designed to gauge the success probability (μ) when only the number of experiments ($n_{\mathcal{S}}$) and the number of successes ($\tau_{\mathcal{S}}$) are known.

In other words, these CIs estimate the **range of "likely" values** the **binomial proportion can take**.

Binomial CIs: Overcoming Wald Limitations

To overcome/reduce the limitations of the Wald interval, Binomial CIs can be adopted.

These CIs are designed to gauge the success probability (μ) when only the number of experiments ($n_{\mathcal{S}}$) and the number of successes ($\tau_{\mathcal{S}}$) are known.

In other words, these CIs estimate the **range of "likely" values** the **binomial proportion can take**.

Proposed solution: the Wilson interval

Also obtained by inverting the acceptance region of the Wald normal test but using the **null standard error** instead of the estimated standard error:

$$\frac{\hat{\mu} + \frac{z^2_{\alpha/2}}{2n_{\mathcal{S}}}}{1 + \frac{z^2_{\alpha/2}}{n_{\mathcal{S}}}} \pm \frac{z_{\alpha/2}}{1 + \frac{z^2_{\alpha/2}}{n_{\mathcal{S}}}} \cdot \sqrt{\frac{\hat{\mu}(1-\hat{\mu})}{n_{\mathcal{S}}} + \frac{z^2_{\alpha/2}}{4n^2_{\mathcal{S}}}}$$

Wilson. Probable Inference, the Law of Succession, and Statistical Inference. J. Amer. Statist. Assoc. 1927

Proposed solution: the Wilson interval

Also obtained by inverting the acceptance region of the Wald normal test but using the **null standard error** instead of the estimated standard error:

Relocated center estimate

$$\frac{\hat{\mu} + \frac{z^2_{\alpha/2}}{2n_{\mathcal{S}}}}{1 + \frac{z^2_{\alpha/2}}{n_{\mathcal{S}}}} \pm \frac{z_{\alpha/2}}{1 + \frac{z^2_{\alpha/2}}{n_{\mathcal{S}}}} \cdot \sqrt{\frac{\hat{\mu}(1-\hat{\mu})}{n_{\mathcal{S}}} + \frac{z^2_{\alpha/2}}{4n^2_{\mathcal{S}}}}$$

Corrected standard deviation

Wilson. Probable Inference, the Law of Succession, and Statistical Inference. J. Amer. Statist. Assoc. 1927

Proposed solution: the Wilson interval

Also obtained by inverting the acceptance region of the Wald normal test but using the **null standard error** instead of the estimated standard error:

Relocated center estimate

$$\frac{\hat{\mu} + \frac{z^2_{\alpha/2}}{2n_{\mathcal{S}}}}{1 + \frac{z^2_{\alpha/2}}{n_{\mathcal{S}}}} \pm \frac{z_{\alpha/2}}{1 + \frac{z^2_{\alpha/2}}{n_{\mathcal{S}}}} \cdot \sqrt{\frac{\hat{\mu}(1-\hat{\mu})}{n_{\mathcal{S}}} + \frac{z^2_{\alpha/2}}{4n^2_{\mathcal{S}}}}$$

Corrected standard deviation

Wilson interval **solves zero-width** and **overshooting issues**
Provides **better coverage** than Wald interval

CAVEAT: Wilson assume SRS as the underlying sampling strategy
Requires design effect adjustments to be used with TWCS and STWCS

Wilson. Probable Inference, the Law of Succession, and Statistical Inference. J. Amer. Statist. Assoc. 1927

Experimental Analysis: Setup

Datasets:

	YAGO	NELL	DisGeNET	SYN 100M
Num. of facts	1,386	1,860	2,999,087	101,415,011
Num. of clusters	822	817	21,243	5,000,000
Avg. cluster size	1.69	2.28	141.18	20.28
Accuracy (μ)	0.99	0.91	n/a	n/a

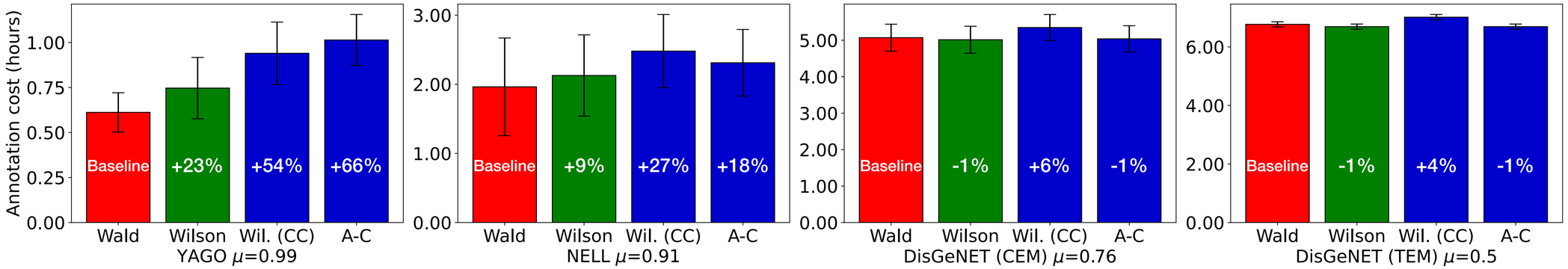
Sampling and Accuracy Estimators:

SRS
TWCS
STWCS

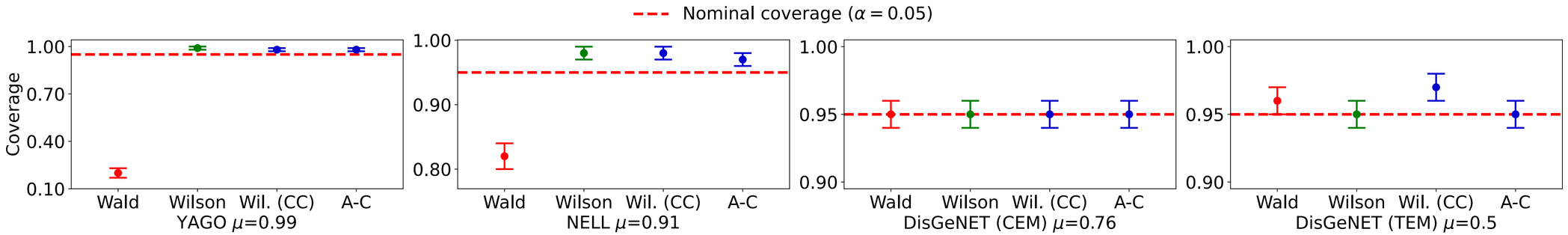
Confidence Interval Estimation:

Wald
Wilson
Continuity-Corrected Wilson
Agresti-Coull

Experimental Analysis: Interval Comparison



(a) Annotation cost comparison under SRS



(b) Coverage comparison under SRS

Wilson most **efficient** solution among **reliable** ones

Wilson best trade-off between efficiency and reliability

Experimental Analysis: Comparison at Scale

	NELL (small)		DisGeNET TEM (medium)		SYN 100M TEM (large)	
	μ = 0.91		μ = 0.90		μ = 0.90	
Method	Cost (hr)	Coverage	Cost (hr)	Coverage	Cost (hr)	Coverage
SRS (Wald)	1.96 ± 0.71	0.82 ± 0.02	2.32 ± 0.77	0.83 ± 0.02	2.38 ± 0.83	0.83 ± 0.02
SRS (Wilson)	2.13 ± 0.59	0.98 ± 0.01	2.46 ± 0.63	0.94 ± 0.02	2.56 ± 0.66	0.93 ± 0.02
TWCS (Wald)	1.56 ± 0.85	0.75 ± 0.03	1.12 ± 0.47	0.82 ± 0.02	1.16 ± 0.50	0.83 ± 0.02
TWCS (Wilson)	1.60 ± 0.82	0.98 ± 0.01	1.14 ± 0.44	0.92 ± 0.02	1.16 ± 0.48	0.93 ± 0.02
STWCS (Wald)	1.32 ± 0.74	0.73 ± 0.03	1.13 ± 0.51	0.78 ± 0.03	1.11 ± 0.57	0.77 ± 0.02
STWCS (Wilson)	1.26 ± 0.72	0.92 ± 0.02	1.18 ± 0.48	0.91 ± 0.02	1.15 ± 0.54	0.88 ± 0.02

Methods **performance** remain **consistent across KGs** with different sizes and topologies
Wilson preserves properties when combined with more advanced sampling techniques

Case Study

[2] Marchesin *et al.* Utility-Oriented Knowledge Graph Accuracy Estimation with Limited Annotations: A Case Study on DBpedia. AAAI HCOMP 2024 (Best Paper Honorable Mention 🎖️)

Utility: Are all Facts Equal?

Different parts of the KG may have **varying utility** depending on **downstream task**

- For entity-oriented search, **popular entities** carry the **highest query load**
  - **Prioritizing** their **assessment** can have a significant impact on the **search experience**

Utility: Are all Facts Equal?

Different parts of the KG may have **varying utility** depending on **downstream task**

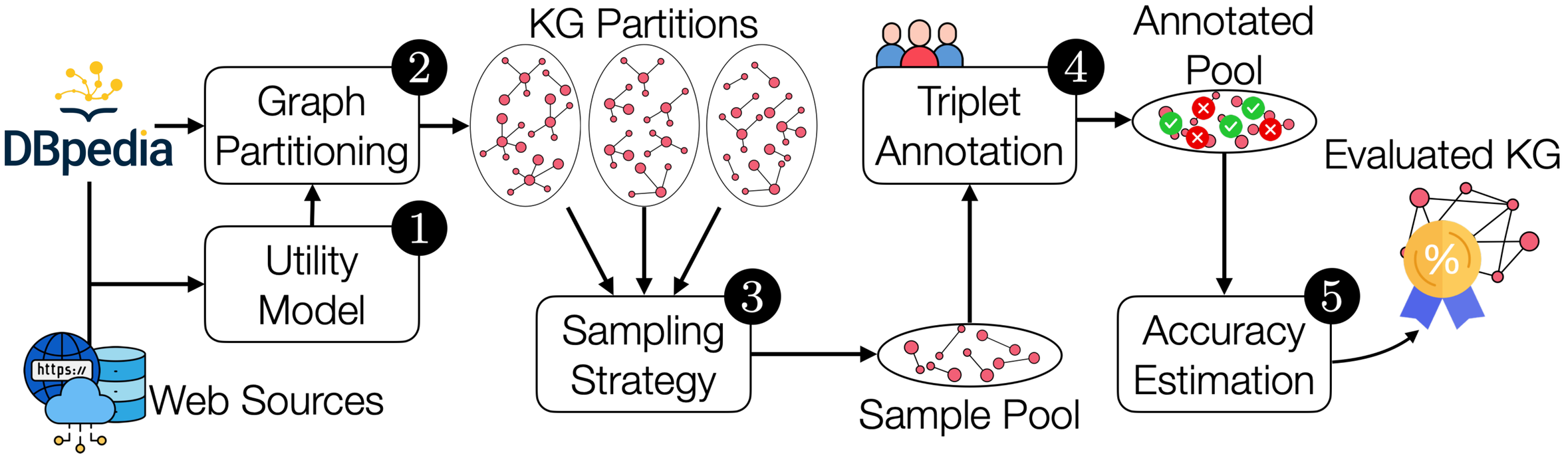
- For entity-oriented search, **popular entities** carry the **highest query load**
  - **Prioritizing** their **assessment** can have a significant impact on the **search experience**

Utility can also help when dealing with **limited annotation budgets**

- Guides **resource allocation** to maximize impact for the downstream task
- Helps decide when to apply **filtering** or **correction mechanisms** in low-quality cases

Utility-Oriented KG Evaluation: A DBpedia Study

DBpedia 2015-10 (6.2M entities and 1.1B triplets)



Utility Model: Popularity in SPARQL Queries (1/2)

We determine the **utility of facts** by their **SPARQL query frequency**

A fact is used in a query either if it appears in the result set or contributes to computing the result

Utility Model: Popularity in SPARQL Queries (1/2)

We determine the **utility of facts** by their **SPARQL query frequency**

A fact is used in a query either if it appears in the result set or contributes to computing the result

- For facts in the result set: execute the queries against the KG

Utility Model: Popularity in SPARQL Queries (1/2)

We determine the **utility of facts** by their **SPARQL query frequency**

A fact is used in a query either if it appears in the result set or contributes to computing the result

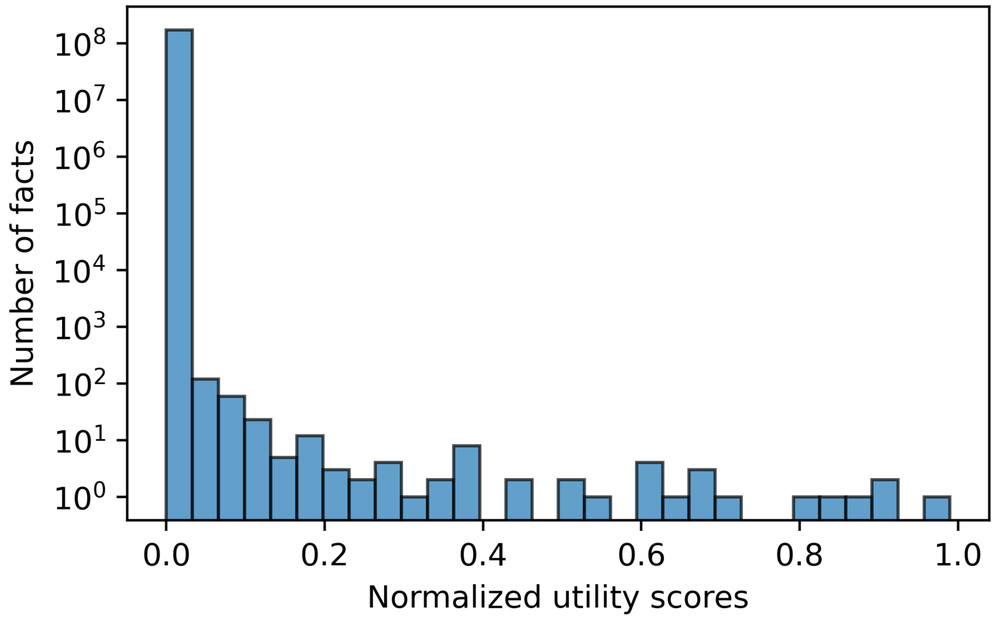
- For facts in the result set: execute the queries against the KG
- For computational facts: compute **query provenance** via lineage
  - Lineage refers to the KG facts that contribute to generating an output fact
  - Transform SPARQL queries into **CONSTRUCT queries**, which retrieve all facts involved in generating the query responses

Dosso et al. Credit distribution in relational scientific databases. Information Systems, 2022

Utility Model: Popularity in SPARQL Queries (2/2)

We used the **LSQ 2.0 dataset** for DBpedia 2015-10 query logs:

- **1.67M SELECT** queries (98%)
- 31K CONSTRUCT, 25 ASK, and 6 DESCRIBE queries (2%)



Most facts cluster near zero, but there is a notable **long tail of high utility facts**

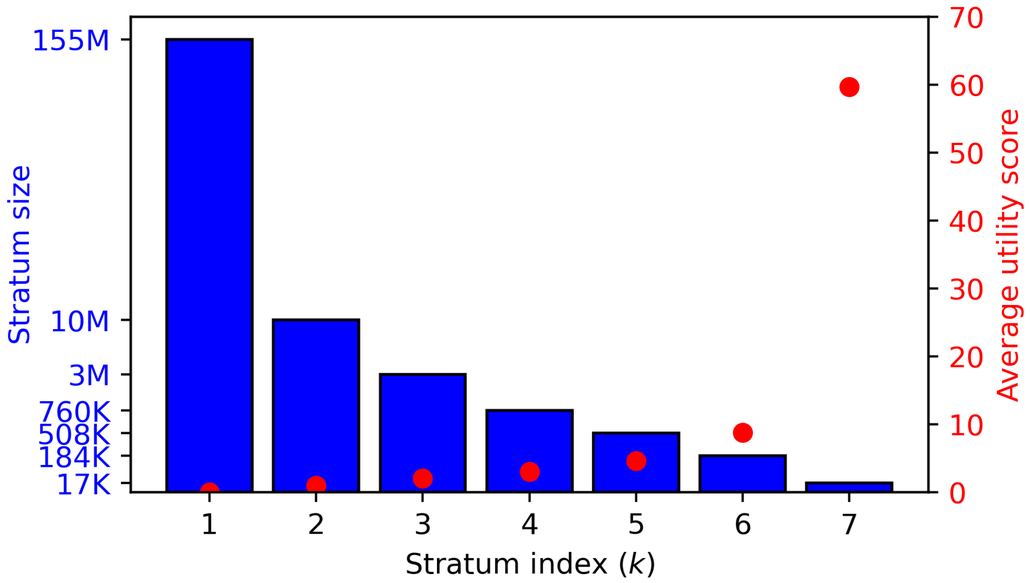
Stadler et al. LSQ 2.0: A linked dataset of SPARQL query logs. Semantic Web, 2024

Graph Partitioning: Stratifying Facts by Utility Scores

We partition the KG into **subsets of facts** based on utility scores

Stratification is done using **Cumulative Square Root of Frequency (CSRF)**

- CSRF minimizes **intra-stratum variance** in utility scores



The number of partitions ($k = 7$) strikes a **balance** between **capturing diverse aspects** and **keeping a reasonable granularity**

Large stratum with **near-zero utility scores** and several **smaller** strata with **higher utility scores**

Dalenius and Hodges. Minimum Variance Stratification. Journal of the American Statistical Association, 1959

KG Accuracy Estimation: Partition- and KG-Level

Fact annotations are used to estimate partition and KG accuracies

Partition accuracy:

$$\hat{\mu}_i = \frac{1}{n_i}\sum_{j=1}^{n_i}\hat{\mu}_{ij},\ \mathrm{Var}(\hat{\mu}_i) = \frac{1}{n_i(n_i-1)}\sum_{j=1}^{n_i}(\hat{\mu}_{ij}-\hat{\mu}_i)^2,\ \mathrm{CI} = \mathrm{Wilson}$$

KG Accuracy Estimation: Partition- and KG-Level

Fact annotations are used to estimate partition and KG accuracies

Partition accuracy:

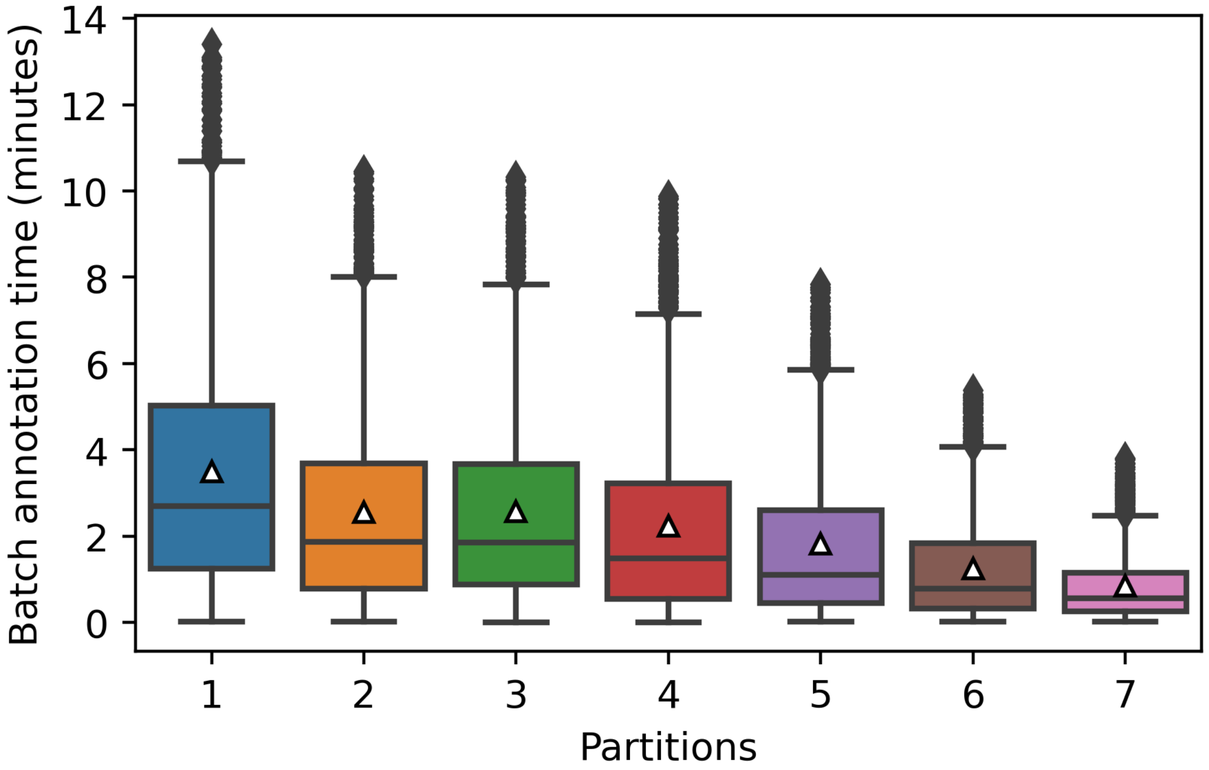
$$\hat{\mu}_i = \frac{1}{n_i} \sum_{j=1}^{n_i} \hat{\mu}_{ij},\ \mathrm{Var}(\hat{\mu}_i) = \frac{1}{n_i(n_i-1)} \sum_{j=1}^{n_i} (\hat{\mu}_{ij} - \hat{\mu}_i)^2,\ \mathrm{CI} = \mathrm{Wilson}$$

KG accuracy:

$$\hat{\mu} = \sum_{i=1}^{k} W_i\, \hat{\mu}_i,\ \mathrm{Var}(\hat{\mu}) = \sum_{i=1}^{k} W_i^2\, \mathrm{Var}(\hat{\mu}_i),\ \mathrm{CI} = \mathrm{Wilson}$$

Experimental Analysis: Annotation Process

Batch annotation time per partition



Annotating batches from **low-utility** partitions demands **more time** than high-utility ones

Given the different costs, the framework gains further importance to **strategic resource allocation**

Experimental Analysis: Statistics & Estimates

37,546 annotations across 9,930 facts from 60 annotators

	Correct	Incorrect	IDK	Total	Estimate
P1	1,851	376	200	2,427	0.83±0.02
P2	1,416	281	110	1,807	0.85±0.02
P3	1,343	246	105	1,694	0.85±0.02
P4	1,164	164	60	1,388	0.89±0.02
P5	927	133	74	1,134	0.89±0.02
P6	715	99	22	836	0.90±0.03
P7	533	96	15	644	0.87±0.03
KG	7,949	1,395	586	9,930	0.83±0.02

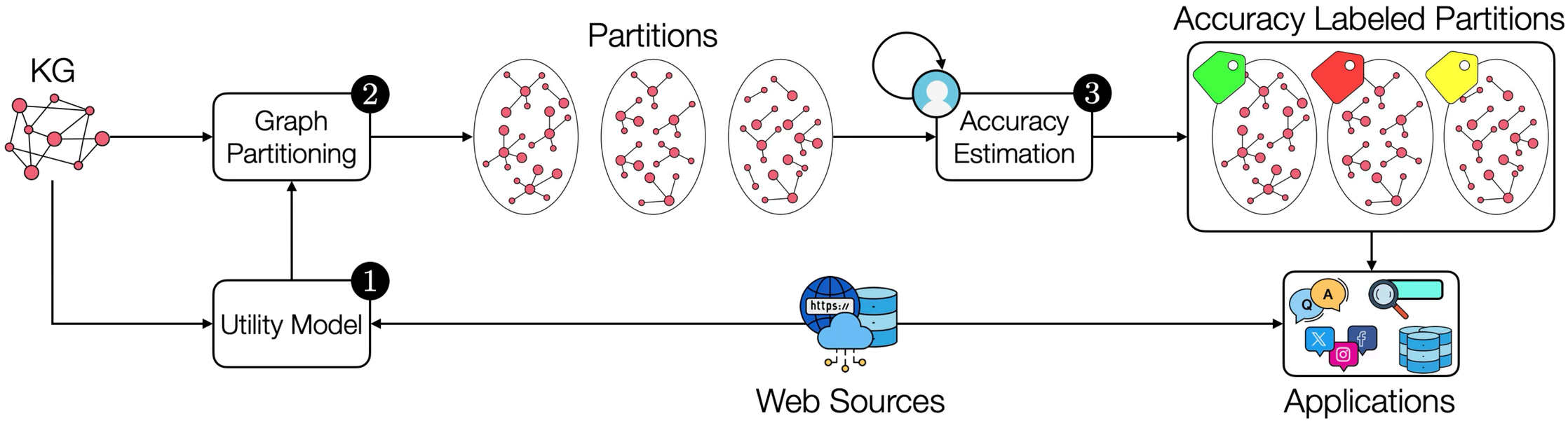
DBpedia 2015-10 is of **high quality**

Estimates have small CIs, providing **strong statistical guarantees**

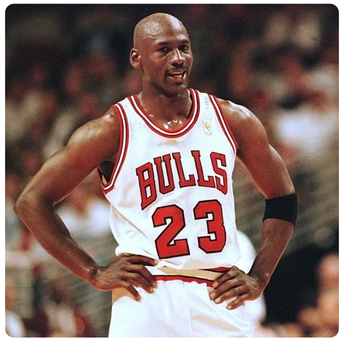
Downstream Application

[3] Marchesin *et al.* Veracity Estimation for Entity-Oriented Search with Knowledge Graphs. CIKM 2024

Quality Estimation for Entity-Oriented Applications



Entity Cards: the Role of Entity Summarization



About

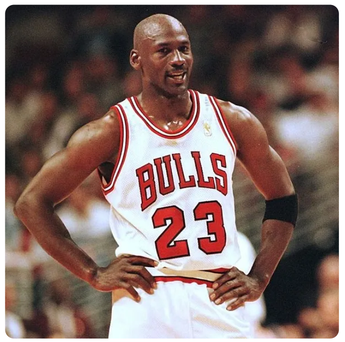
Michael Jeffrey Jordan, also known by his initials MJ, is an American businessman and former professional basketball player. His profile on the official National Basketball Association website states that "by acclamation, Michael Jordan is the greatest basketball player of all time." Wikipedia

Born: February 17, 1963, Cumberland Hospital
Net worth: 3 billion USD (2023) Forbes
Spouse: Yvette Prieto (m. 2013), Juanita Vanoy (m. 1989–2006)
Children: Marcus Jordan, Jeffrey Michael Jordan, Jasmine M. Jordan
Number: 23 (Chicago Bulls / Shooting guard), MORE
Teammates: Scottie Pippen, Magic Johnson, Larry Bird, MORE
Parents: James R. Jordan, Sr., Deloris Jordan

Extracting the most **important facts** about an **entity**

Addressed as a **ranking problem**: rank facts by importance and **pick top-*k*** ones

Entity Cards: the Role of Entity Summarization

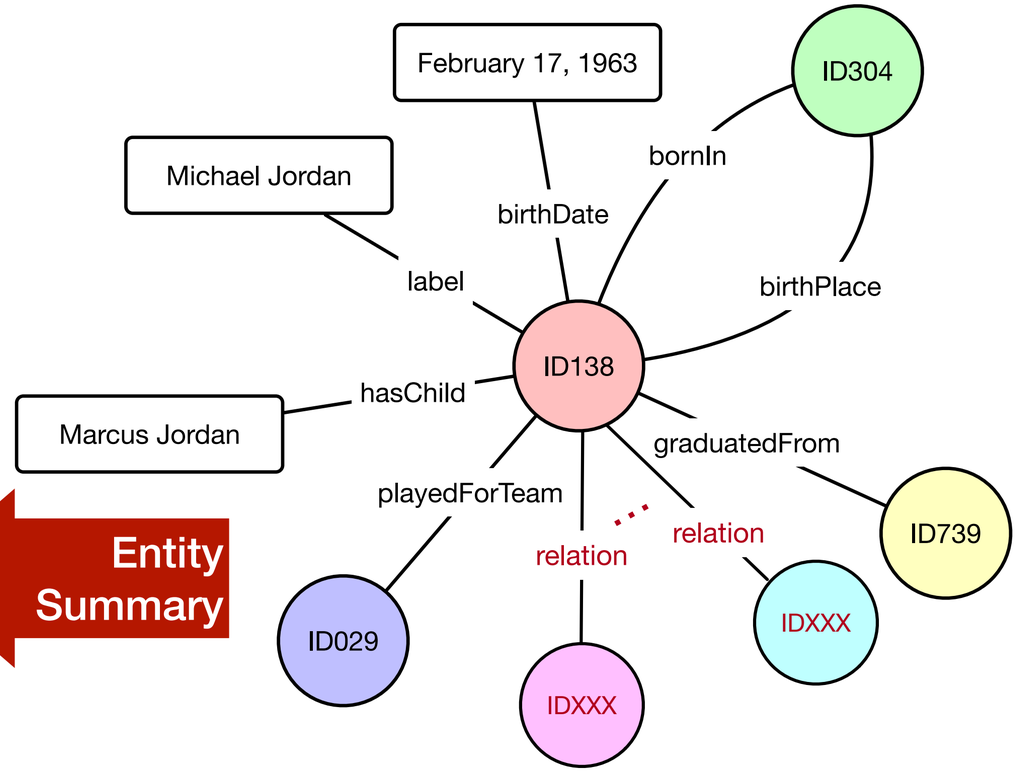


About

Michael Jeffrey Jordan, also known by his initials MJ, is an American businessman and former professional basketball player. His profile on the official National Basketball Association website states that "by acclamation, Michael Jordan is the greatest basketball player of all time." Wikipedia

Born: February 17, 1963, Cumberland Hospital
Net worth: 3 billion USD (2023) Forbes
Spouse: Yvette Prieto (m. 2013), Juanita Vanoy (m. 1989–2006)
Children: Marcus Jordan, Jeffrey Michael Jordan, Jasmine M. Jordan
Number: 23 (Chicago Bulls / Shooting guard), MORE
Teammates: Scottie Pippen, Magic Johnson, Larry Bird, MORE
Parents: James R. Jordan, Sr., Deloris Jordan

Entity Summary



Extracting the most **important facts** about an **entity**

Addressed as a **ranking problem**: rank facts by importance and **pick top-*k*** ones

Dynamic Entity Summarization

How to generate **query-dependent entity summaries** that can directly address **users' information needs**?

Ranking entity **facts based** on their **importance** for the entity and **relevance** to the query

Hasibi et al. Dynamic Factual Summaries for Entity Cards. SIGIR, 2017

Dynamic Entity Summarization

How to generate **query-dependent entity summaries** that can directly address **users' information needs**?

Ranking entity **facts based** on their **importance** for the entity and **relevance** to the query

Born: March 14, 1879, Ulm, Germany
Died: April 18, 1955, Princeton, New Jersey, United States
Siblings: Maja Einstein
Spouse: Elsa Einstein (m. 1919-1936), Mileva Marić (m. 1903-1919)
Education: University of Zurich (1905), more

Hasibi et al. Dynamic Factual Summaries for Entity Cards. SIGIR, 2017

Dynamic Entity Summarization

How to generate **query-dependent entity summaries** that can directly address **users' information needs**?

Ranking entity **facts based** on their **importance** for the entity and **relevance** to the query

Einstein family

Born: March 14, 1879, Ulm, Germany
Died: April 18, 1955, Princeton, New Jersey, United States
Siblings: Maja Einstein
Spouse: Elsa Einstein (m. 1919-1936), Mileva Marić (m. 1903-1919)
Education: University of Zurich (1905), more

Einstein awards

Born: March 14, 1879, Ulm, Germany
Died: April 18, 1955, Princeton, New Jersey, United States
Education: University of Zurich (1905), more
Awards: Barnard Medal for Meritorious Service to Science, Academy Award, more
Influenced: Satyendra Nath Bose, Wolfgang Pauli, Tom Cruise, more

Hasibi et al. Dynamic Factual Summaries for Entity Cards. SIGIR, 2017

Dynamic Entity Summarization

How to generate **query-dependent entity summaries** that can directly address **users' information needs**?

Ranking entity **facts based** on their **importance** for the entity and **relevance** to the query

Einstein family

Born: March 14, 1879, Ulm, Germany
Died: April 18, 1955, Princeton, New Jersey, United States
Siblings: Maja Einstein
Spouse: Elsa Einstein (m. 1919-1936), Mileva Marić (m. 1903-1919)
Education: University of Zurich (1905), more

Einstein awards

Born: March 14, 1879, Ulm, Germany
Died: April 18, 1955, Princeton, New Jersey, United States
Education: University of Zurich (1905), more
Awards: Barnard Medal for Meritorious Service to Science, Academy Award, more
Influenced: Satyendra Nath Bose, Wolfgang Pauli, Tom Cruise, more

Relevance and importance do NOT consider the correctness (veracity) of facts!

Hasibi et al. Dynamic Factual Summaries for Entity Cards. SIGIR, 2017

Dynamic Entity Summarization

How to generate **query-dependent entity summaries** that can directly address **users' information needs**?

Ranking entity **facts based** on their **importance** for the entity and **relevance** to the query

Einstein family

Born: March 14, 1879, Ulm, Germany
Died: April 18, 1955, Princeton, New Jersey, United States
Siblings: Maja Einstein
Spouse: Elsa Einstein (m. 1919-1936), Mileva Marić (m. 1903-1919)
Education: University of Zurich (1905), more

Einstein awards

Born: March 14, 1879, Ulm, Germany
Died: April 18, 1955, Princeton, New Jersey, United States
Education: University of Zurich (1905), more
Awards: Barnard Medal for Meritorious Service to Science, Academy Award, more
Influenced: Satyendra Nath Bose, Wolfgang Pauli, Tom Cruise, more

Relevance and importance do NOT consider the correctness (veracity) of facts!

How to increase entity cards reliability? **Prioritize** facts w/ **higher** (estimated) **quality**

Hasibi et al. Dynamic Factual Summaries for Entity Cards. SIGIR, 2017

Experimental Analysis: Setup

RQ: What is the impact of KG quality on entity search systems?

Experimental Analysis: Setup

RQ: What is the impact of KG quality on entity search systems?

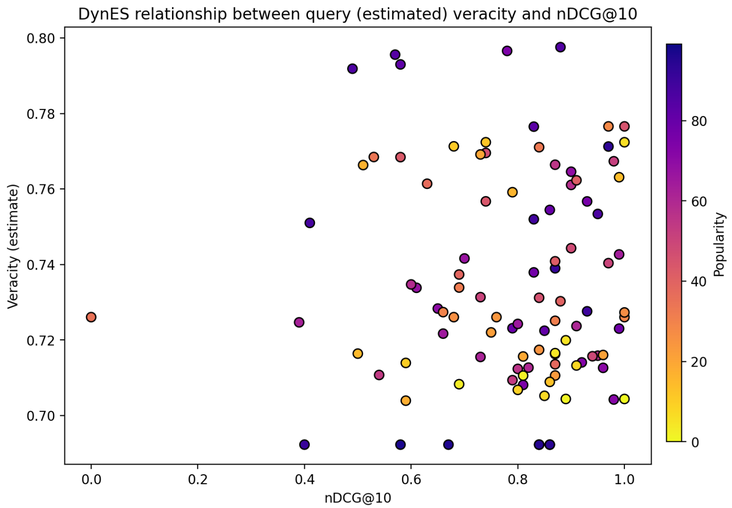
Collection: Dynamic entity summarization collection based on **DBpedia 2015-10**

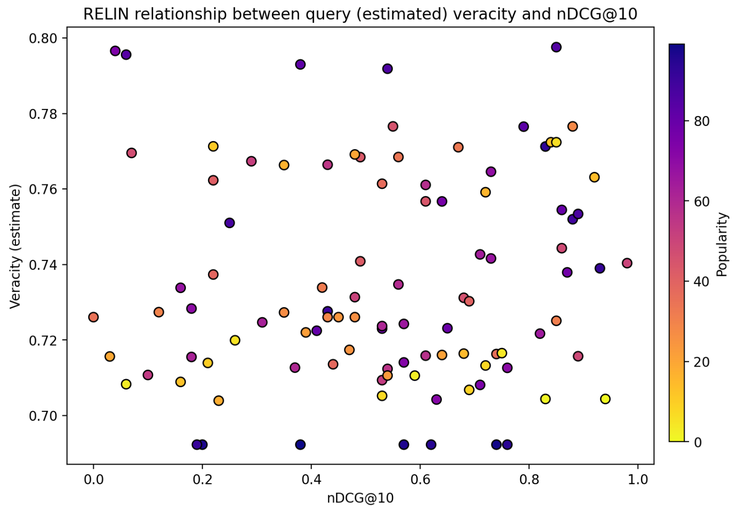
Summarization methods (SOTA):

- DynES - learning-to-rank method that generates query-dependent entity summaries
- RELIN - PageRank based method selecting top-*h* facts based on relatedness and informativeness

Veracity re-ranking (*v*Rank): prioritize facts w/ higher (estimated) quality

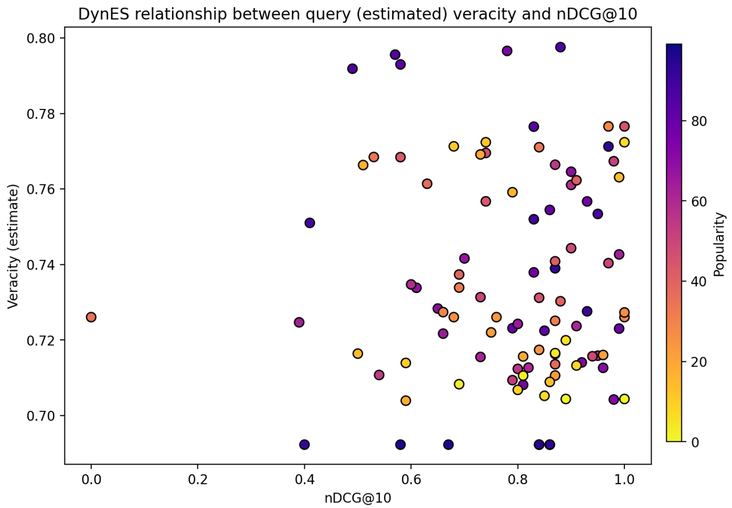
Experimental Analysis: Accuracy Impact on Ranking

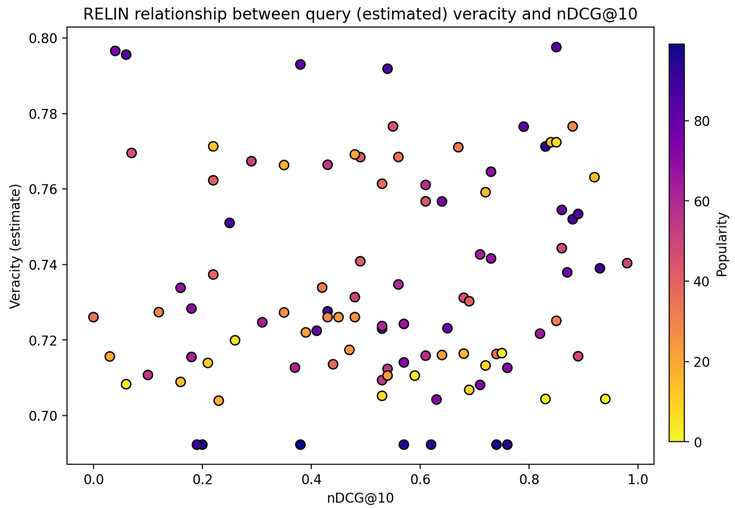




Accuracy and relevance are **orthogonal**

Experimental Analysis: Accuracy Impact on Ranking





	nDCG@5	nDCG@10
DynES (orig)	0.76	0.79
DynES (*v*Rank)	0.76	0.79
RELIN (orig)	0.46	0.52
RELIN (*v*Rank)	0.46	0.53

Accuracy and relevance are **orthogonal**

*v*Rank strategy **boosts quality** while **keeping retrieval effectiveness**

Experimental Analysis: Accuracy Impact on Entity Cards

Query: Directed Bela Glen Glenda Bride Monster Plan 9 Outer Space

Entity Cards:

(A) Bride of the Monster

country History of the United States (1945–64)
producer Ed Wood
starring Bela Lugosi
director Ed Wood
writer Ed Wood

(B) Bride of the Monster

country History of the United States (1945–64)
starring Bela Lugosi
starring Loretta King Hadler
producer Ed Wood
director Ed Wood

- A is better
- B is better
- They are the same

Submit Annotations

Five expert annotators:

- *v*Rank better in 29%
- *v*Rank inferior in 23%
- *v*Rank equivalent in 48%

*v*Rank either proves **better** or **maintains** user perception w/o detriment in **77%**

Conclusions

Conclusions: the Importance of KG Quality

- **Quality matters:** reliable KGs are foundational for trustworthy AI and information access systems
- **Limitations exist** in traditional evaluation methods — prompting a need for more robust and efficient approaches
- **Scalable frameworks** enable broader, deeper assessments of KG quality
- **Utility-focused evaluation** connects accuracy to real-world use and impact

Future Work: the Road Ahead 

- **Bayesian approaches** for more informative and appropriate post-data inference (accepted at SIGMOD 2025)
- Reliable methods to efficiently update quality estimates as **KGs evolve over time**
- Use of **ontologies** and **inference rules** to further **reduce human involvement**

Marchesin and Silvello. Credible Intervals for Knowledge Graph Accuracy Estimation. SIGMOD 2025

ANY Questions?

References

- Marchesin and Silvello. Efficient and Reliable Estimation of Knowledge Graph Accuracy. VLDB 2024
- Marchesin *et al*. Utility-Oriented Knowledge Graph Accuracy Estimation with Limited Annotations: A Case Study on DBpedia. AAAI HCOMP 2024
- Marchesin *et al*. Veracity Estimation for Entity-Oriented Search with Knowledge Graphs. CIKM 2024
- Marchesin and Silvello. Credible Intervals for Knowledge Graph Accuracy Estimation. SIGMOD 2025