Data Science with Graphs

– Exploration –

Matteo Lissandrini – University of Verona

Outline

1. **Intro to Graph Exploration**
   - Taxonomy of Graph Exploration
   - KG profiling & summarization
   - Exploratory Search
   - Example Based Exploration
2. **Node-based Exploratory Search**
   - Seed-set expansion
   - Minimum Wiener Connector problem
   - Focused Clustering
   - Entity Set Search
3. **Structure-based Exploratory Search**
   - Reverse-engineering Queries
   - Entity Tuples
   - Exemplar Queries
   - Example-Based Graph suggestion

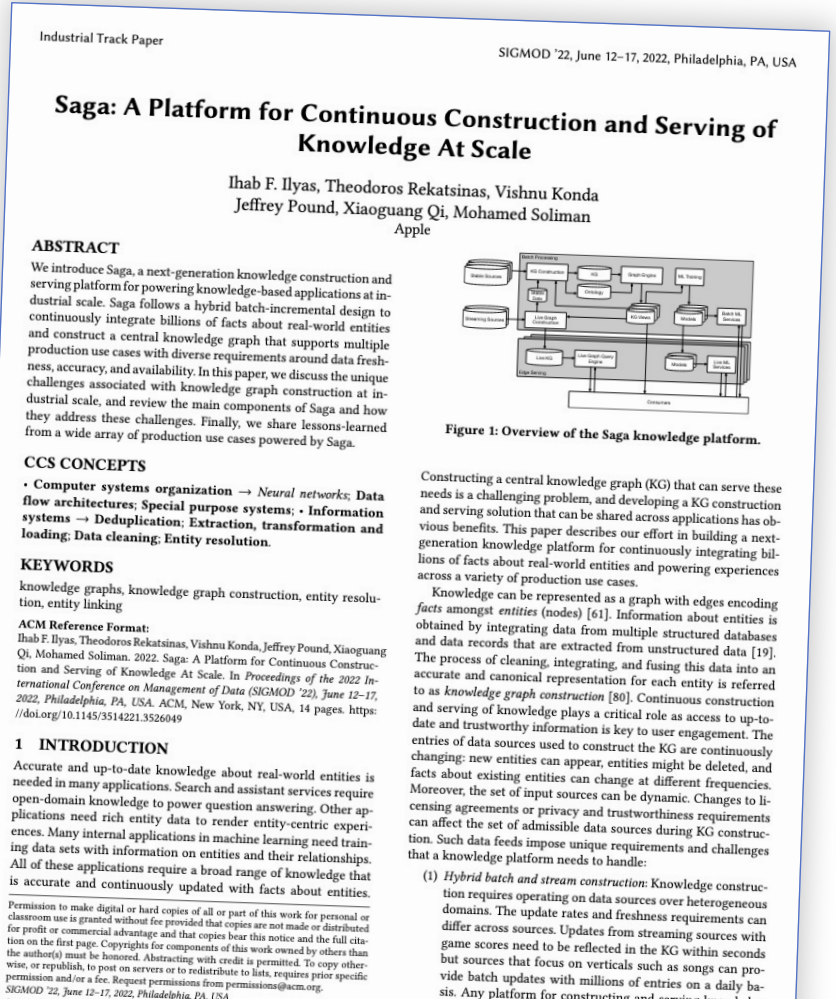
We are all "Data Novices" at some point...

The entries of data sources used to construct the KG **are continuously changing...**

[...]

Self-serve data onboarding: Low-effort **onboarding of new data sources** is important to ensure consistent growth of the KG.

The Data Novice:
A user unfamiliar with the data at hand and its structure



Industrial Track Paper SIGMOD '22, June 12–17, 2022, Philadelphia, PA, USA

Saga: A Platform for Continuous Construction and Serving of Knowledge At Scale

Ihab F. Ilyas, Theodoros Rekatsinas, Vishnu Konda
Jeffrey Pound, Xiaoguang Qi, Mohamed Soliman
Apple

ABSTRACT

We introduce Saga, a next-generation knowledge construction and serving platform for powering knowledge-based applications at industrial scale. Saga follows a hybrid batch-incremental design to continuously integrate billions of facts about real-world entities and construct a central knowledge graph that supports multiple production use cases with diverse requirements around data freshness, accuracy, and availability. In this paper, we discuss the unique challenges associated with knowledge graph construction at industrial scale, and review the main components of Saga and how they address these challenges. Finally, we share lessons-learned from a wide array of production use cases powered by Saga.

CCS CONCEPTS

• **Computer systems organization** → *Neural networks*; **Data flow architectures**; **Special purpose systems**; • **Information systems** → **Deduplication**; **Extraction, transformation and loading**; **Data cleaning**; **Entity resolution**.

KEYWORDS

knowledge graphs, knowledge graph construction, entity resolution, entity linking

ACM Reference Format:
Ihab F. Ilyas, Theodoros Rekatsinas, Vishnu Konda, Jeffrey Pound, Xiaoguang Qi, Mohamed Soliman. 2022. Saga: A Platform for Continuous Construction and Serving of Knowledge At Scale. In *Proceedings of the 2022 International Conference on Management of Data (SIGMOD '22), June 12–17, 2022, Philadelphia, PA, USA.* ACM, New York, NY, USA, 14 pages. https://doi.org/10.1145/3514221.3526049

1 INTRODUCTION

Accurate and up-to-date knowledge about real-world entities is needed in many applications. Search and assistant services require open-domain knowledge to power question answering. Other applications need rich entity data to render entity-centric experiences. Many internal applications in machine learning need training data sets with information on entities and their relationships. All of these applications require a broad range of knowledge that is accurate and continuously updated with facts about entities.

SIGMOD '22, June 12–17, 2022, Philadelphia, PA, USA

Figure 1: Overview of the Saga knowledge platform.

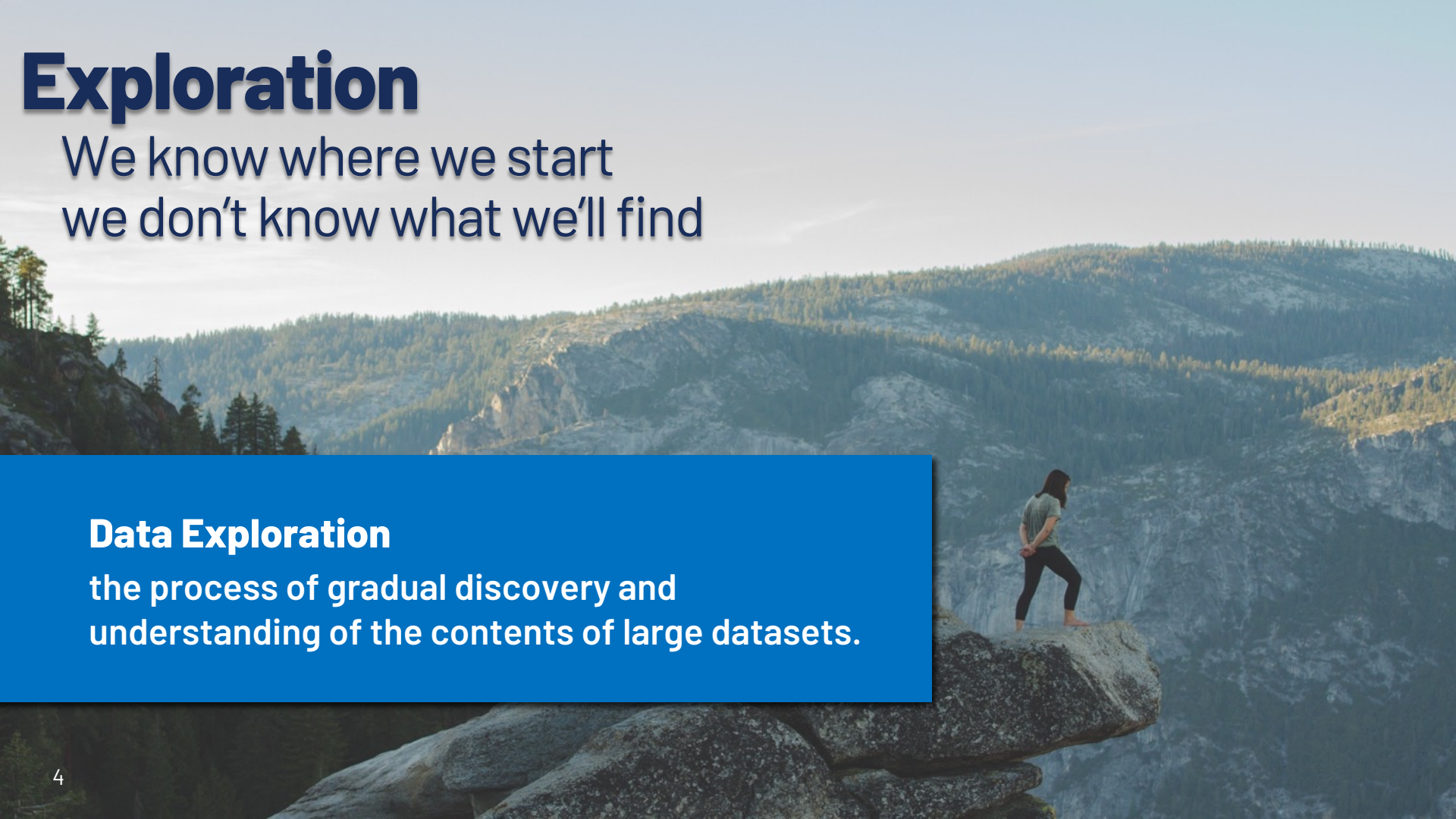
Constructing a central knowledge graph (KG) that can serve these needs is a challenging problem, and developing a KG construction and serving solution that can be shared across applications has obvious benefits. This paper describes our effort in building a next-generation knowledge platform for continuously integrating billions of facts about real-world entities and powering experiences across a variety of production use cases.

Knowledge can be represented as a graph with edges encoding *facts* amongst *entities* (nodes) [61]. Information about entities is obtained by integrating data from multiple structured databases and data records that are extracted from unstructured data [19]. The process of cleaning, integrating, and fusing this data into an accurate and canonical representation for each entity is referred to as *knowledge graph construction* [80]. Continuous construction and serving of knowledge plays a critical role as access to up-to-date and trustworthy information is key to user engagement. The entries of data sources used to construct the KG are continuously changing: new entities can appear, entities might be deleted, and facts about existing entities can change at different frequencies. Moreover, the set of input sources can be dynamic. Changes to licensing agreements or privacy and trustworthiness requirements can affect the set of admissible data sources during KG construction. Such data feeds impose unique requirements and challenges that a knowledge platform needs to handle:

(1) *Hybrid batch and stream construction:* Knowledge construction requires operating on data sources over heterogeneous domains. The update rates and freshness requirements can differ across sources. Updates from streaming sources with game scores need to be reflected in the KG within seconds but sources that focus on verticals such as songs can provide batch updates with millions of entries on a daily basis. Any platform for constructing and

Exploration

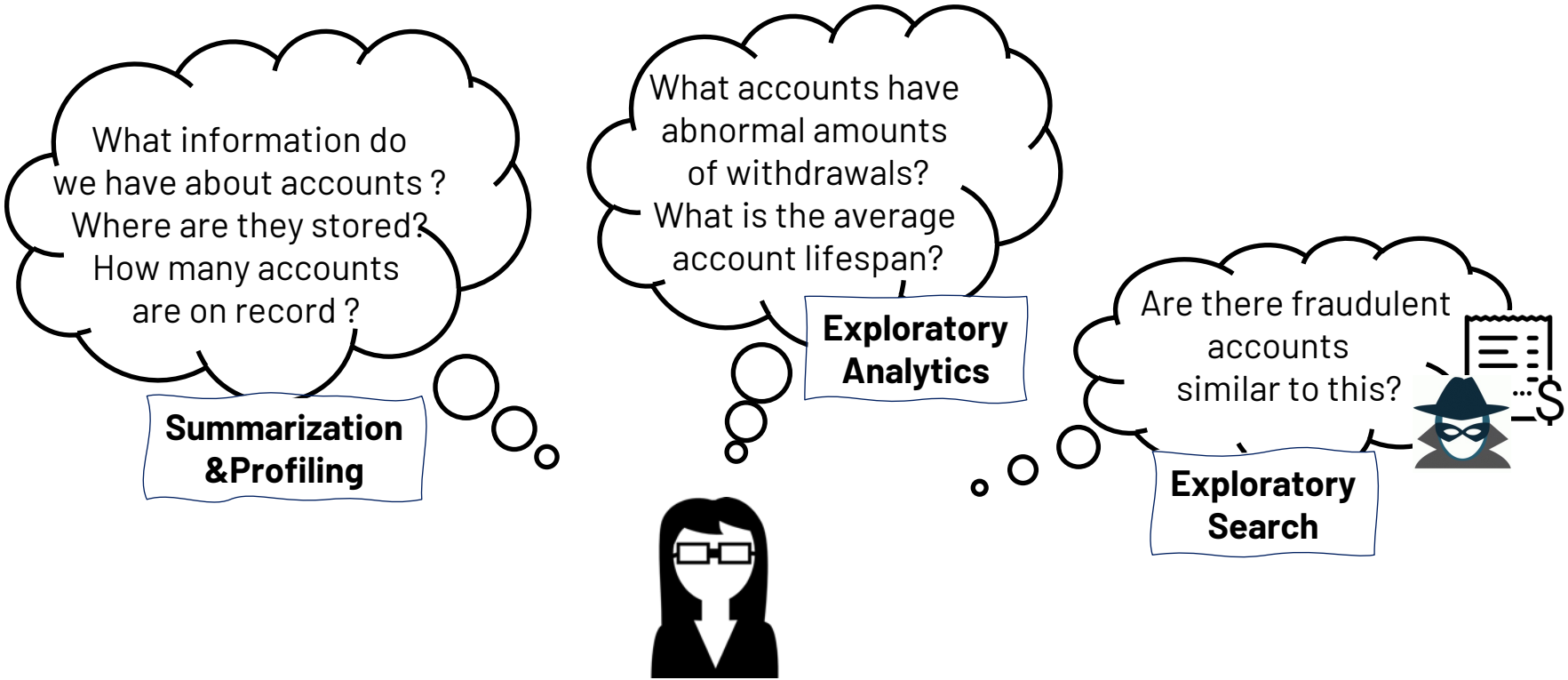
We know where we start
we don't know what we'll find

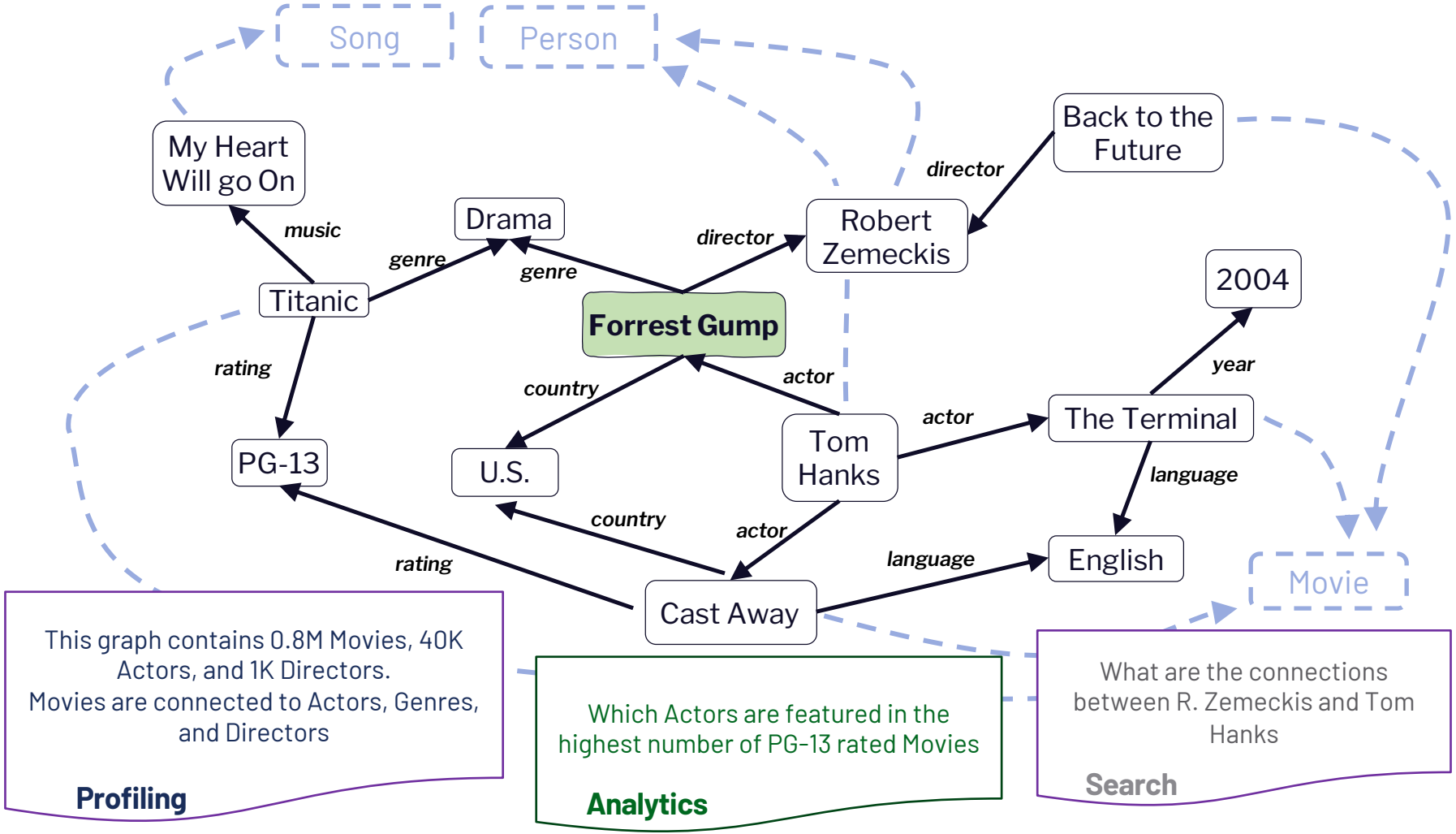


Data Exploration

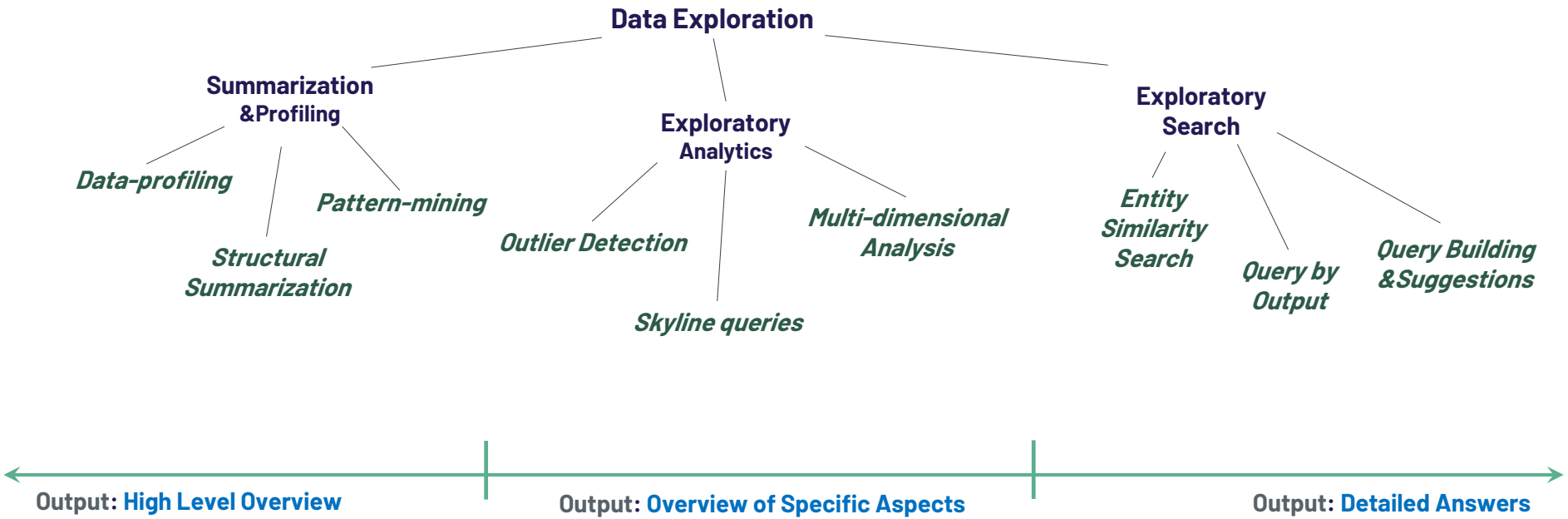
the process of gradual discovery and understanding of the contents of large datasets.

Data Exploration Needs

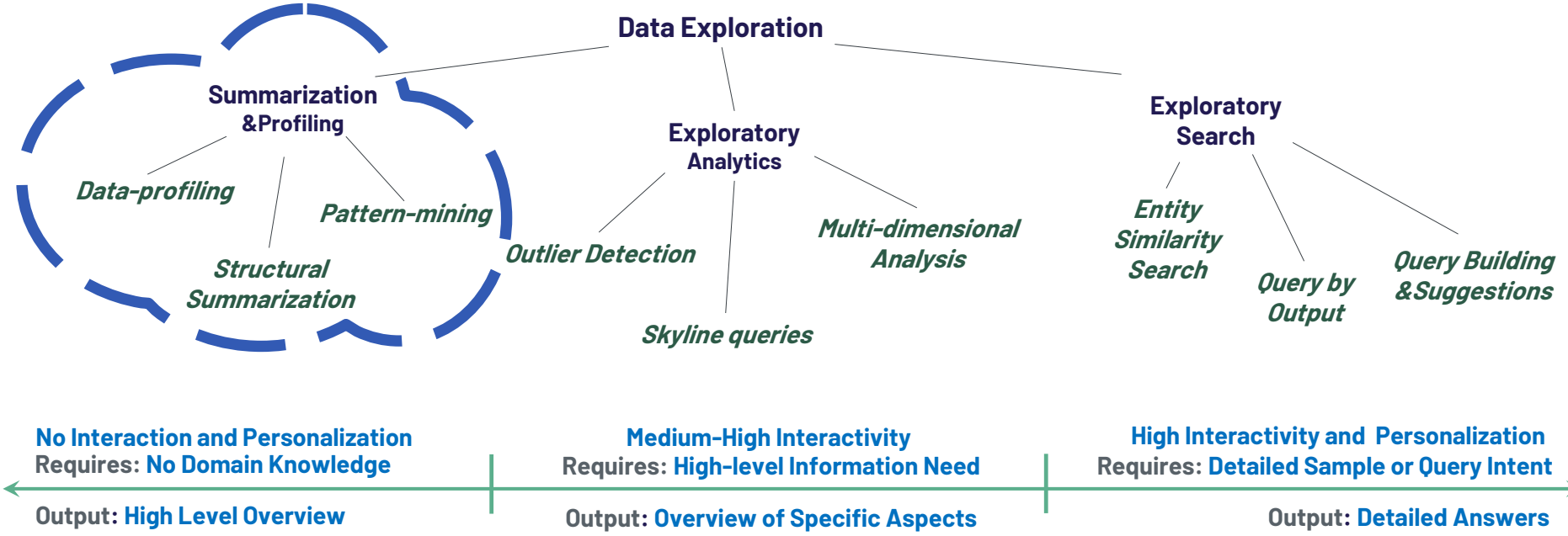



Song
Person
My Heart Will go On
Back to the Future
director
music
Drama
director
Robert Zemeckis
genre
genre
Titanic
Forrest Gump
2004
rating
year
country
actor
actor
The Terminal
PG-13
U.S.
Tom Hanks
language
country
actor
rating
language
English
Movie
Cast Away
This graph contains 0.8M Movies, 40K Actors, and 1K Directors.
Movies are connected to Actors, Genres, and Directors
Profiling
Which Actors are featured in the highest number of PG-13 rated Movies
Analytics
What are the connections between R. Zemeckis and Tom Hanks
Search

Data Exploration Methods



Data Exploration Methods



KG Profiling

Heiko Paulheim [2017]
http://ceur-ws.org/Vol-1927/paper1.pdf

Obtain a basic understanding of the contents of a KG

1. How many instances? How many classes?
2. What's the vocabulary (predicates/attributes)
3. Are there big-hubs? Are there disconnected islands?

Table 1: Global Properties of the Knowledge Graphs compared in this paper

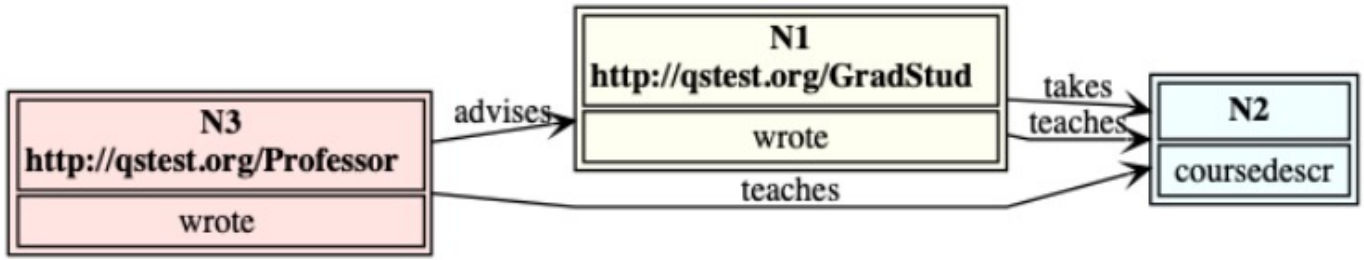
	DBpedia	YAGO	Wikidata	OpenCyc	NELL
Version	2016-04	YAGO3	2016-08-01	2016-09-05	08m.995
# instances	5,109,890	5,130,031	17,581,152	118,125	1,974,297
# axioms	397,831,457	1,435,808,056	1,633,309,138	2,413,894	3,402,971
avg. indegree	13.52	17.44	9.83	10.03	5.33
avg. outdegree	47.55	101.86	41.25	9.23	1.25
# classes	754	576,331	30,765	116,822	290
# relations	3,555	93,659	11,053	165	1,334
Releases	biyearly	> 1 year	live	> 1 year	1-2 days

KG Summarization & Pattern Mining

Surveys:
Kellou-Menouer et al [2022]
Čebirić et al [2019]

Extract overall structural information: quotient graphs

1. How are classes connected?
2. Which predicates and attributes are shared by entities of this type?
3. What is the prevalence of connections across nodes with this properties?



https://link.springer.com/article/10.1007/s00778-021-00717-x
https://link.springer.com/article/10.1007/s00778-018-0528-3

Shape Extraction Analysis

Datasets

Profiling

	DBpedia	LUBM	YAGO-4	WikiData-2015	WikiData-2021
# of triples	52 M	91 M	210 M	290 M	1.926 B
# of objects	19 M	12 M	126 M	64 M	617 M
# of subjects	15 M	10 M	5 M	40 M	196 M
# of literals	15 M	5.5 M	111 M	40 M	904 M
# of instances	5 M	1 M	17 M	3 M	91 M
# of classes	427	22	8,902	13,227	82,693
# of properties	1,323	20	153	4,906	9,017
Size in GBs	6.6	15.66	28.59	42	234

Size and characteristics of the datasets used for experiments

M: Million, B: Billion

Shape Extraction Analysis: Extracted Structures

Datasets

Profiling

Table 2: Shapes Statistics using QSE-Exact.

	NS	PS	Non-Literal PSc	Literal PSc
	COUNT	COUNT/AVG	COUNT/AVG	COUNT/AVG
LUBM	23	164 / 7.1	323 / 3.0	57 /1.0
DBpedia	426	11,916 / 27.9	38,454 / 6.9	5,335 /1.0
YAGO-4	8,897	76,765 / 8.6	315,413 / 14.5	50,708 / 1.0
Wdt15	13,227	202,085 / 15.2	114,890 / 3.0	106,599 / 1.0
Wdt21	82,651	2,051,538 / 24.8	3,765,953 / 5.6	1,113,856 / 1.0

NS = Node Sapes
PS = Property Shapes

Why is this important?

Sometimes, even though you think of your data as a graph, it is not really a graph!

LSQB: A Large-Scale Subgraph Query Benchmark

Amine Mhedhbi
University of Waterloo
amine.mhedhbi@uwaterloo.ca

Matteo Lissandrini
Aalborg University
matteo@cs.aau.dk

Laurens Kuiper
CWI Amsterdam
laurens.kuiper@cwi.nl

Jack Waudby
Newcastle University
j.waudby2@newcastle.ac.uk

Gábor Szárnyas
CWI Amsterdam
gabor.szarnyas@cwi.nl

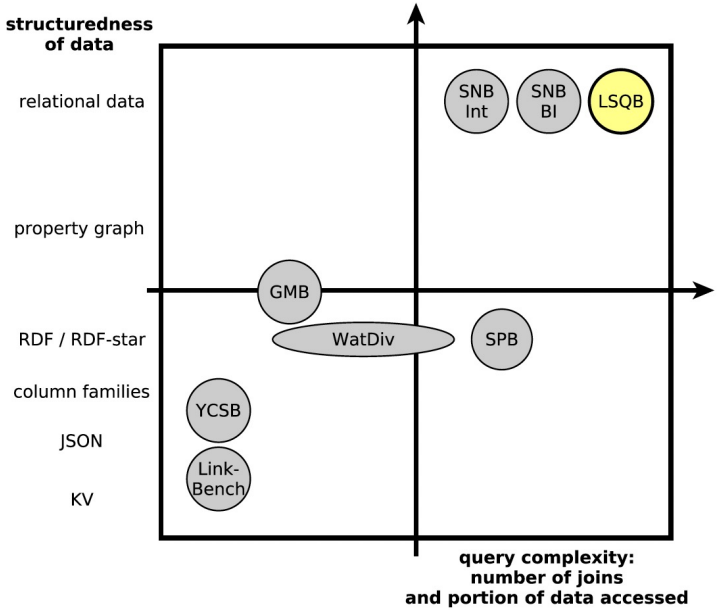
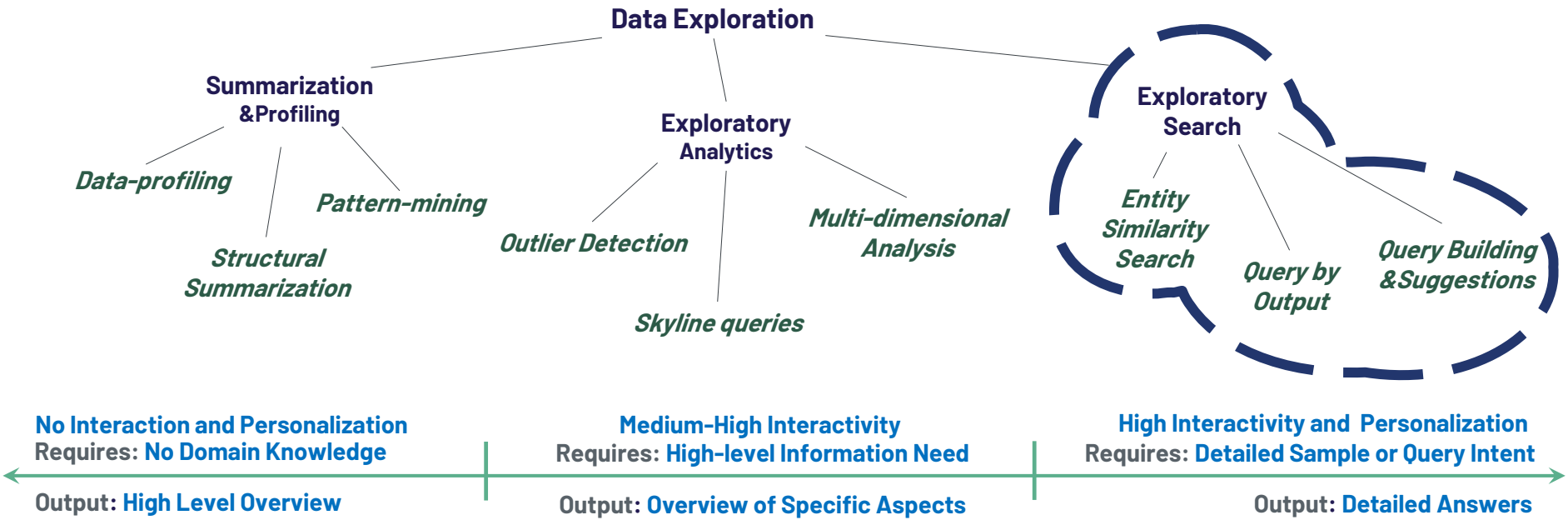
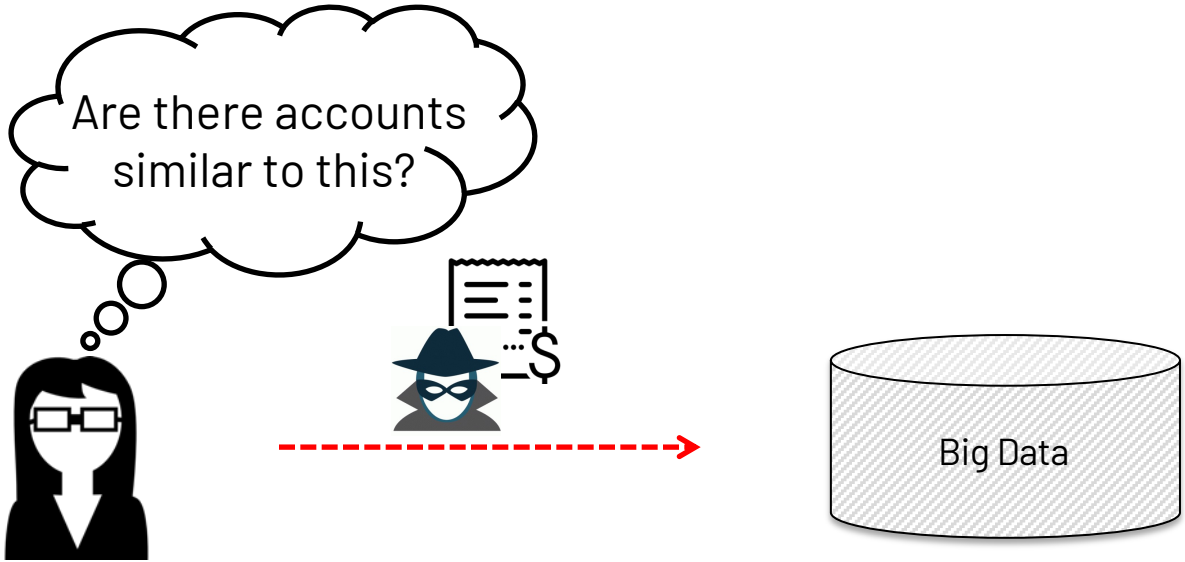


Figure 1: Landscape of DB benchmarks according to query complexity and data structuredness. The benchmark proposed in this paper, LSQB, is highlighted.

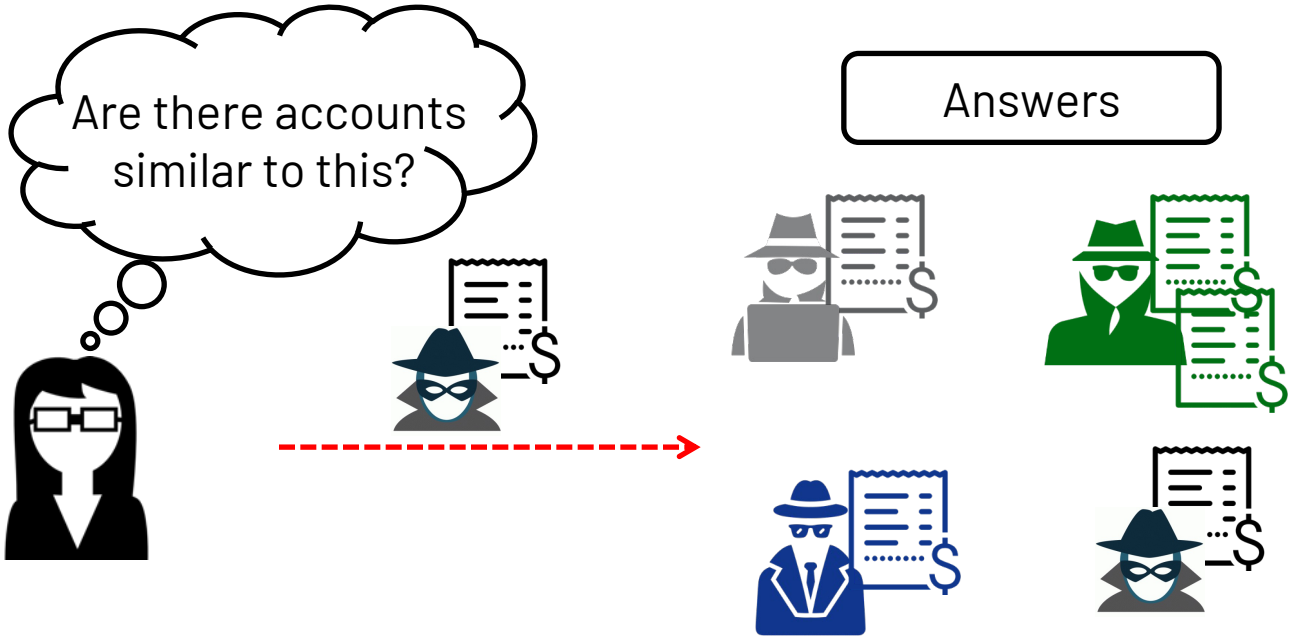
Data Exploration Methods



Examples as Exploratory Methods



Examples as Exploratory Methods



Example is always more efficacious than precept
Samuel Johnson, Rasselas (1759)

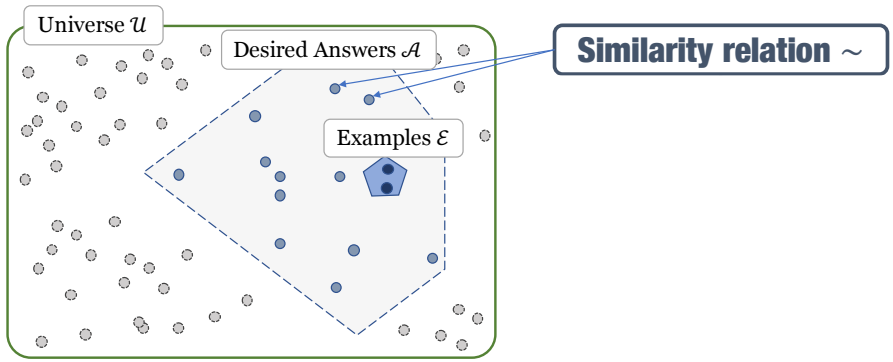
Similarities are the key . . .

If we knew how similar each item is with respect to any other for each user, we would know the answer

The Example-based problem

Given

a set of examples $\mathcal{E}$ from a universe $\mathcal{U}$



Find

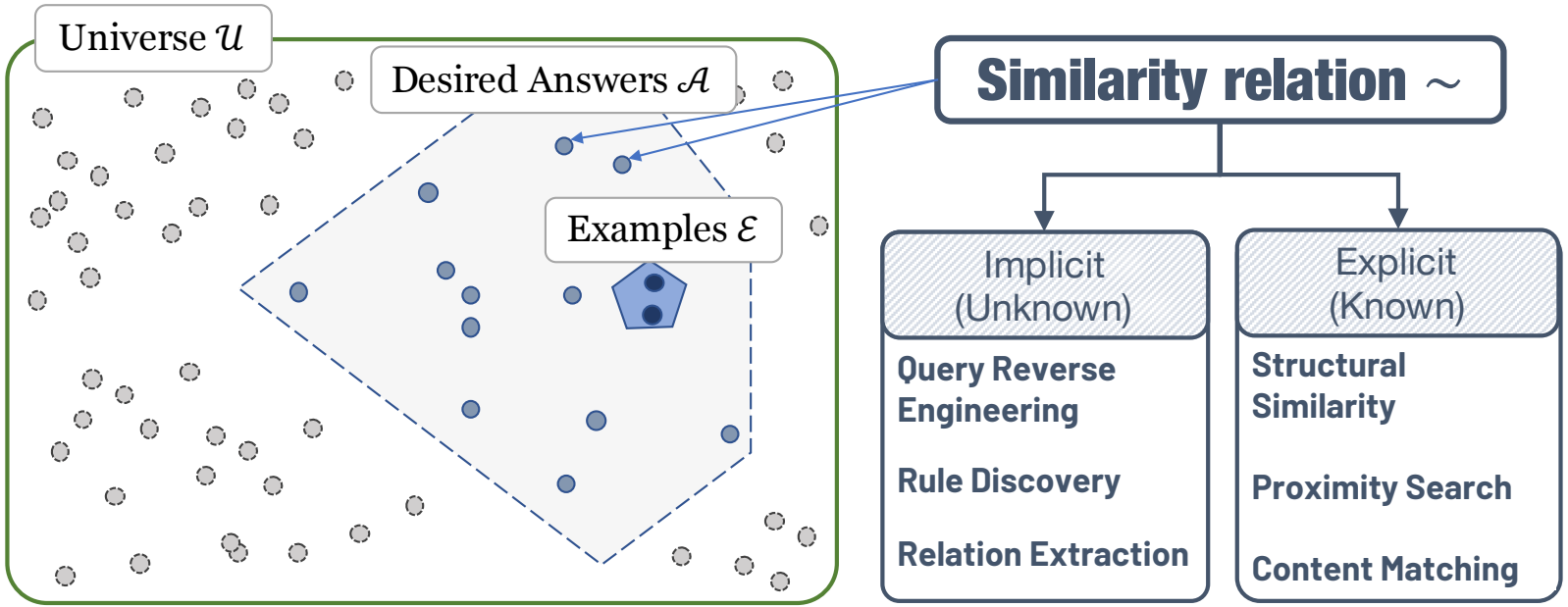
a similarity " $\sim$ "

such that

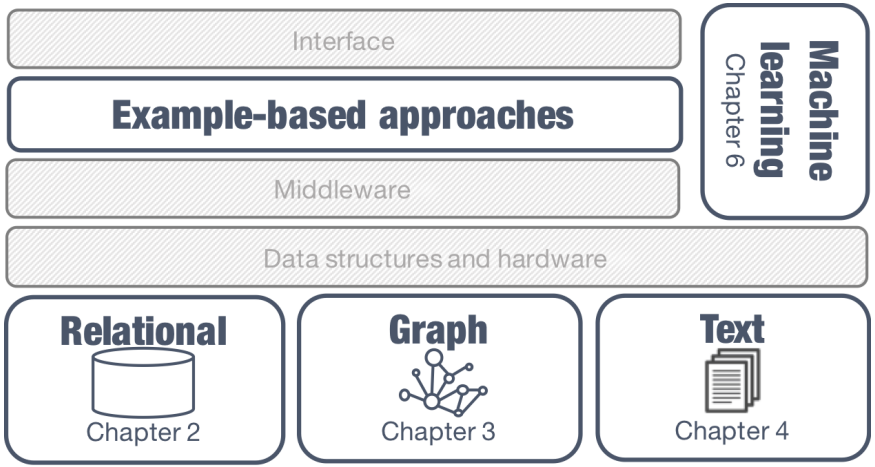
1. When $\mathcal{E}$ is part of the answers $\mathcal{A}$ (partially or totally)
2. The answers in $\mathcal{A}$ are the most similar to the examples in $\mathcal{E}$ according to " $\sim$ "

What similarity " $\sim$ " should we use ?
How do we identify " $\sim$ " (for each user) ?

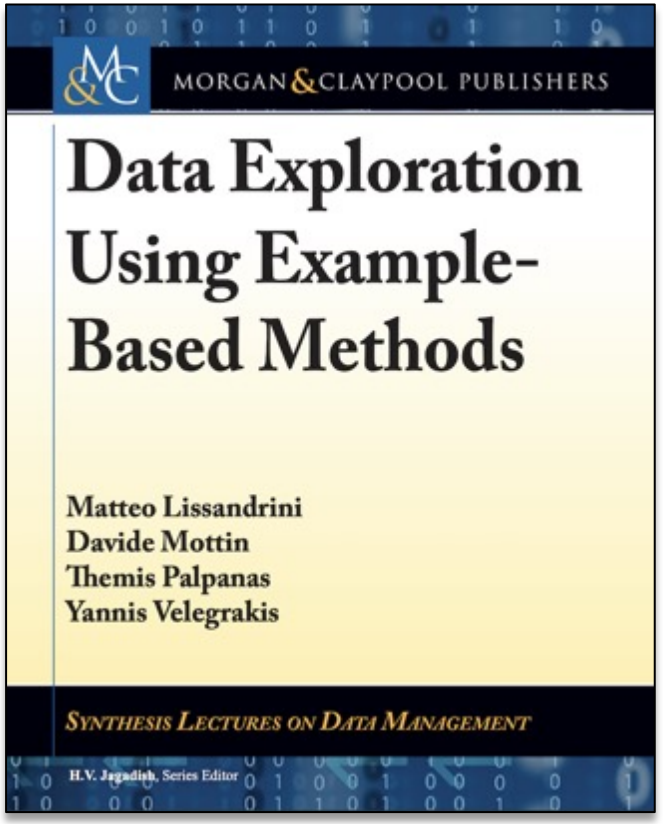
Example-based methods



Book on Example-based methods



Slides and Materials
https://data-exploration.eu/

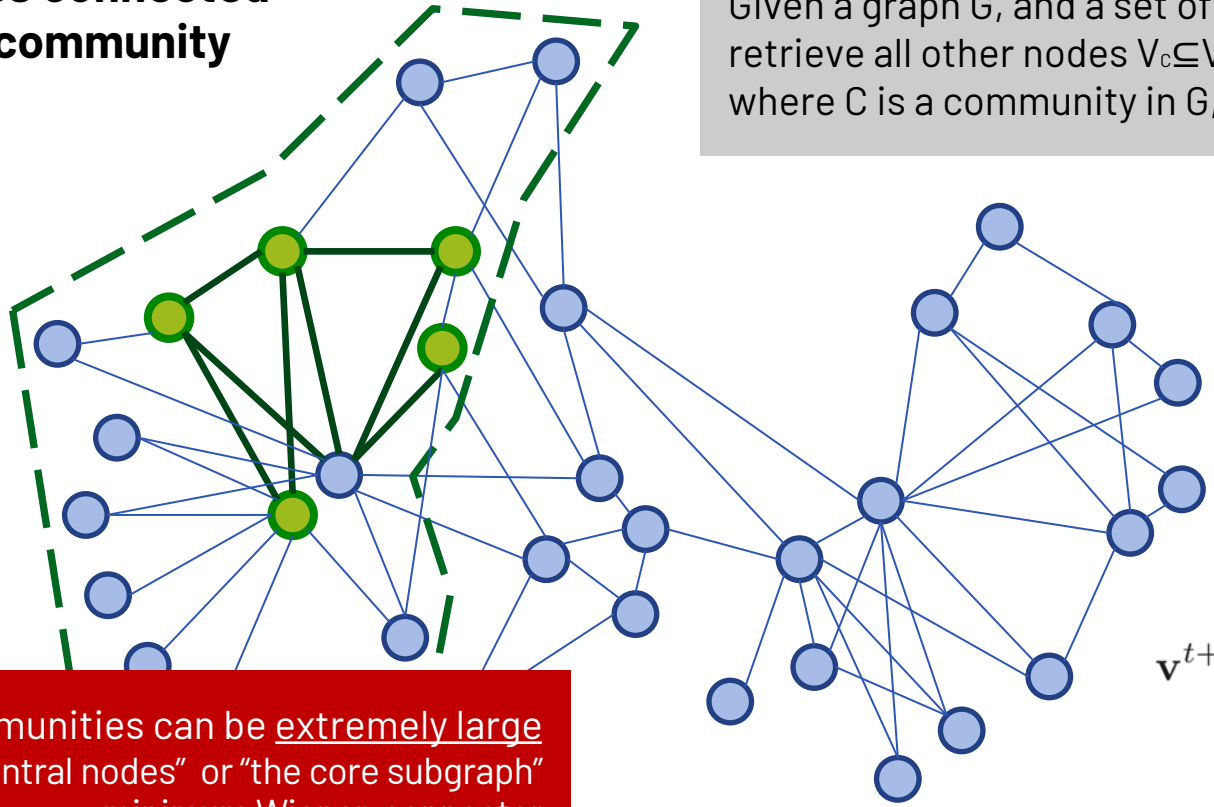


Seed Set Expansion

Nodes connected by a community

Kloumann and Kleinberg [2014]

Given a graph G, and a set of query nodes $V_Q \subseteq V_G$, retrieve all other nodes $V_C \subseteq V_G$, where C is a community in G, and $V_Q \subseteq V_C$.



Solution: PPR

$$\mathbf{v}^{t+1} = (1-\alpha)\mathbf{M} \cdot \mathbf{v}^t + \alpha \mathbf{v}^0$$

Communities can be extremely large
Identify "central nodes" or "the core subgraph"
minimum Wiener-connector

Focused Clustering and Outlier Detection

Similarity based on attributes

Model: Unlabeled Undirected Graph with Node Attributes

Query: A set of Nodes Q

Similarity: To Be Inferred

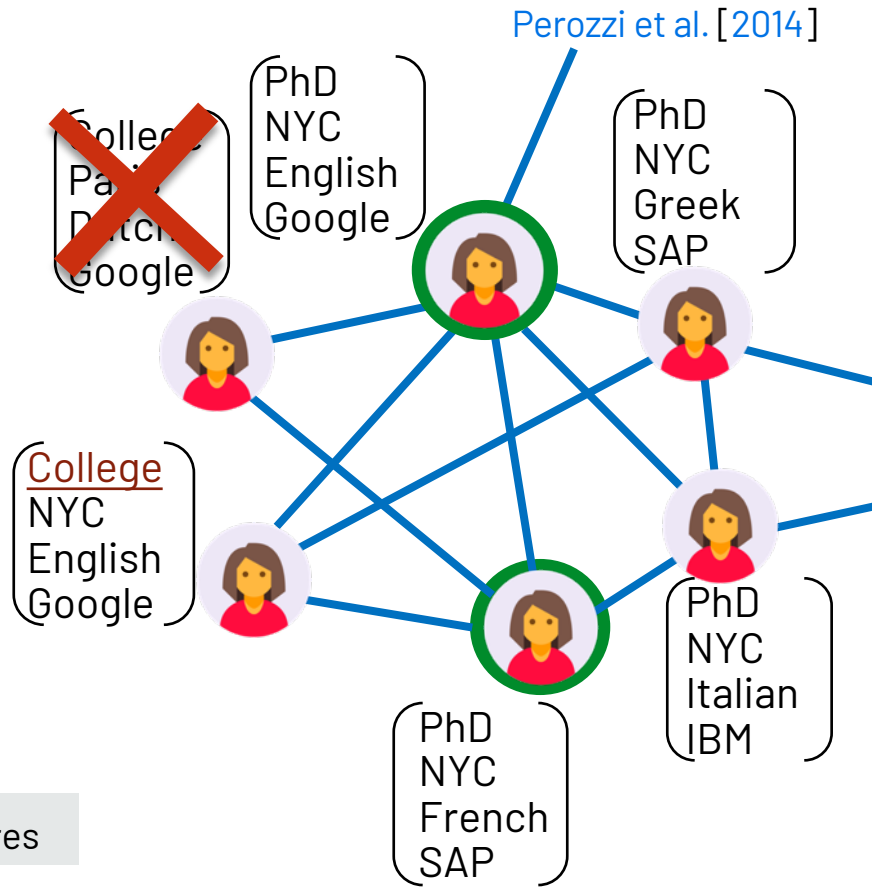
based on Attribute Values & Connectivity

Output: Clusters of Nodes: Dense & Coherent

+ Outliers

Case: Target Users ⟶ Community with same interests

Case: Products⟶ Co-purchased products with similar features



Focused Clustering

Infer User Focus

TASK: Infer "FOCUS" , important attributes

attribute weights β

PhD NYC English Google

PhD NYC French SAP

→ 0.5 0.5 0 0

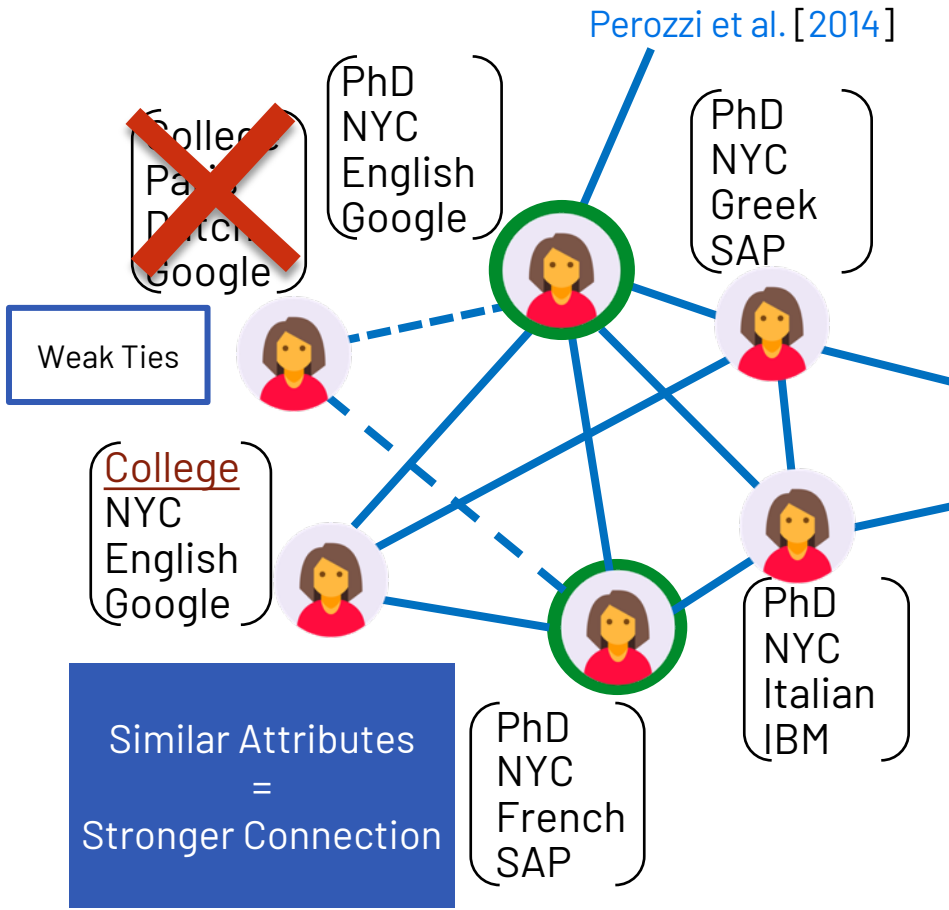
1. Set of similar pairs, PS (from Q)

2. Set of dissimilar pairs, PD (random sample)

3. Learn a distance metric between PS and PD

$$\min_A \sum_{(u,v)\in P_S} (f_i - f_j)^T \mathbf{A} (f_i - f_j) - \gamma \log \left(\sum_{(u,v)\in P_D} \sqrt{(f_i - f_j)^T \mathbf{A} (f_i - f_j)} \right)$$

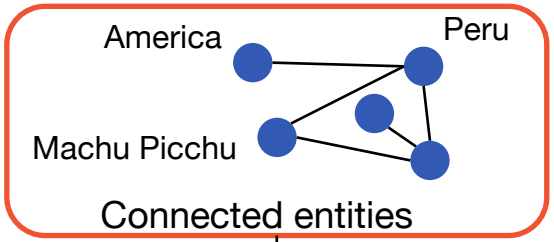
(Distance Metric Learning, inverse Mahalanobis distance: Xing, et al 2002)



Serendipitous Search

Bordino et al. [2013]

Enhance document links with Entities and Query-logs



Input: Query/Document
Output: Queries

Serenditpity
Related topics potentially come to mind *after* consulting the page.

Document

Serendipitous Search

rafting excursion down the urubamba river
el dorado temple of sun
indios quechuas
map of peru
sapa inca

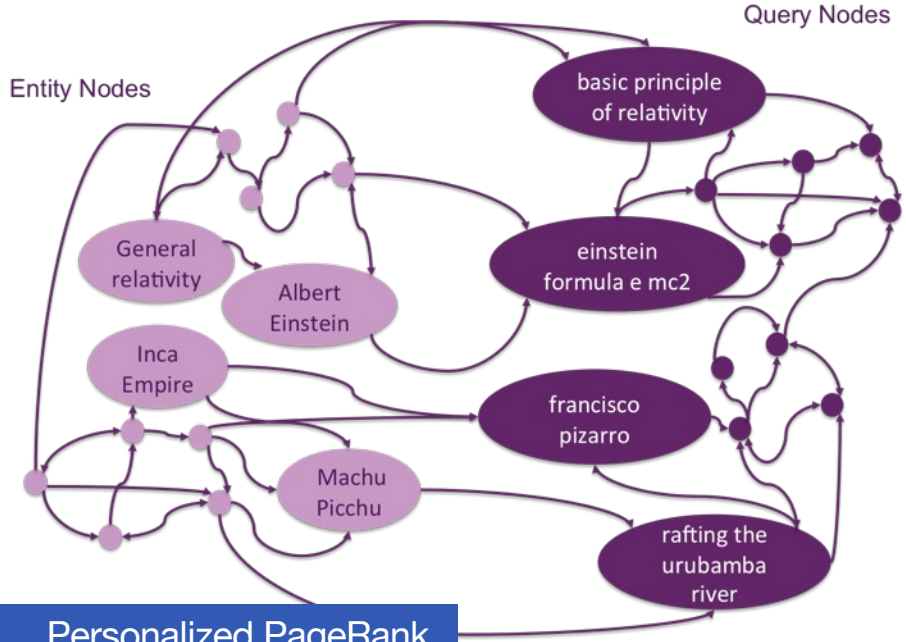
Searches related to Document content

Exploit "lateral connections" in User Search Behaviors

Entity Query Graph

Bordino et al. [2013]

Entity-Query graph from queries to entities and back



Personalized PageRank to score suggested queries

EQGraph Weighted Edges

Queries in the same session

1. query to query:

$$w_Q(q_i \to q_j) = w_{QFG}(q_i \to q_j)$$

Frequency-based approach

2. entity to query

$$w_{EQ}(e \to q) = \frac{f(q)}{\sum_{q_i | e \in X_E(q_i)} f(q_i)}$$

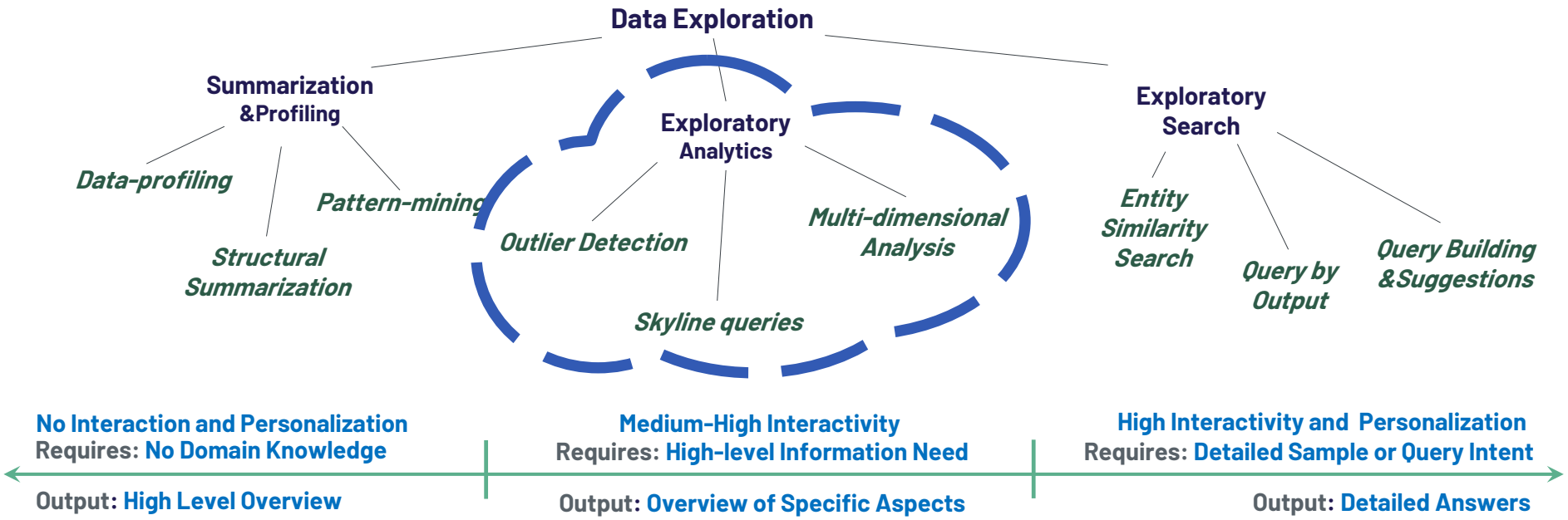
The more queries entities share the higher the probability

3. entity to entity

$$w_E(e_u \to e_v) = 1 - \prod_{i=1,\ldots,r} (1 - p_{q_{i_s} \to q_{i_t}}(e_u \to e_v))$$

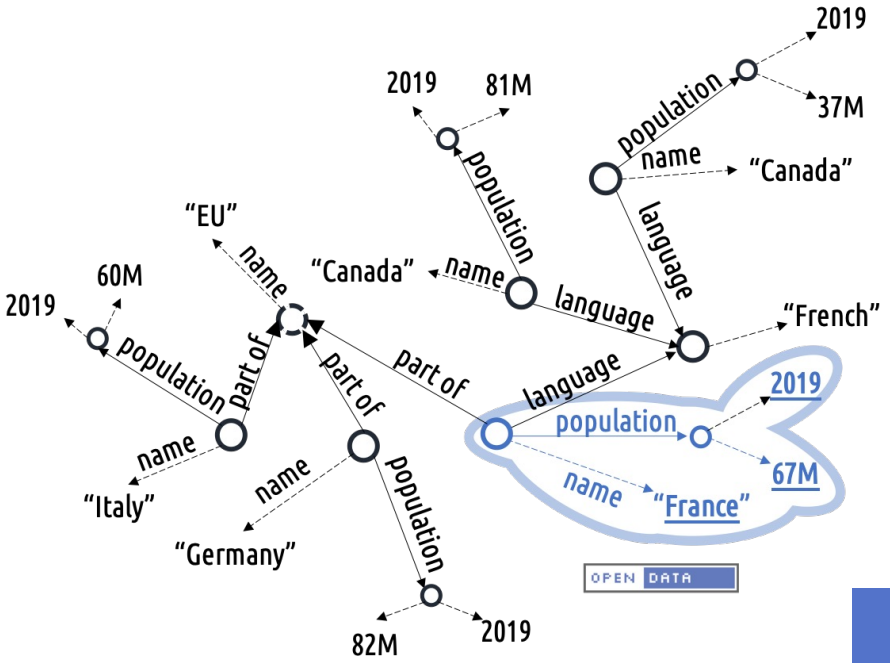
Based on query to query edges

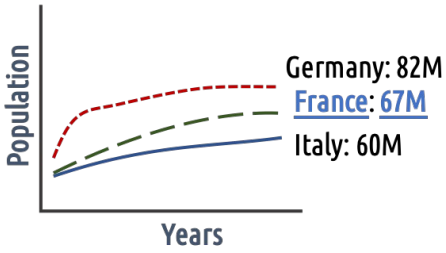
Data Exploration Methods

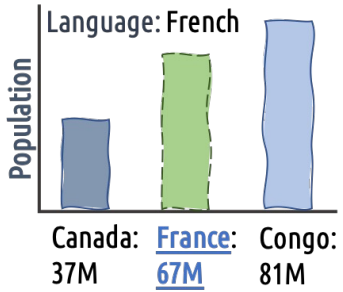


ReOLAP: Reverse Engineering OLAP Queries on Knowledge Graphs

Lissandrini et al.
[EDBT 2023]







APPROACH:
Query Reverse Engineering

What About LLMs?

Can we allow an LLM to manipulate the structure of the graph?

When graphs are associated with text or when the graph contains textual attributes, how can we exploit LLMs ability to manipulate text?

Large Language Models on Graphs: A Comprehensive Survey

Bowen Jin*, Gang Liu*, Chi Han*, Meng Jiang, Heng Ji, Jiawei Han

Abstract—Large language models (LLMs), such as GPT4 and LLaMA, are creating significant advancements in natural language processing, due to their strong text encoding/decoding ability and newly found emergent capability (*e.g.*, reasoning). While LLMs are mainly designed to process pure texts, there are many real-world scenarios where text data is associated with rich structure information in the form of graphs (*e.g.*, academic networks, and e-commerce networks) or scenarios where graph data are paired with rich textual information (*e.g.*, molecules with descriptions). Besides, although LLMs have shown their pure text-based reasoning ability, it is underexplored whether such ability can be generalized to graphs (*i.e.*, graph-based reasoning). In this paper, we provide a systematic review of scenarios and techniques related to large language models on graphs. We first summarize potential scenarios of adopting LLMs on graphs into three categories, namely pure graphs, text-attributed graphs, and text-paired graphs. We then discuss detailed techniques for utilizing LLMs on graphs, including LLM as Predictor, LLM as Encoder, and LLM as Aligner, and compare the advantages and disadvantages of different schools of models. Furthermore, we discuss the real-world applications of such methods and summarize open-source codes and benchmark datasets. Finally, we conclude with potential future research directions in this fast-growing field. **The related source can be found at https://github.com/PeterGriffinJin/Awesome-Language-Model-on-Graphs.**

Index Terms—Large Language Models, Graph Neural Networks, Natural Language Processing, Graph Representation Learning

1 Introduction

Large language models (LLMs) (*e.g.*, BERT [23], T5 [29], LLaMA [119]) which represents a direction of

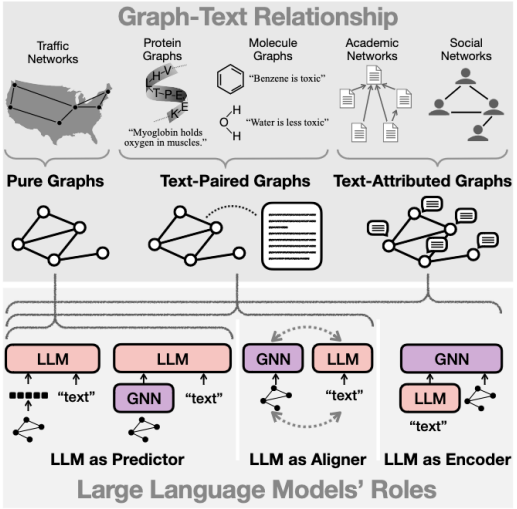


Fig. 1. According to the relationship between graph and text, we categorize three LLM on graph scenarios. Depending on the role of LLM, we summarize three LLM-on-graph techniques. "LLM as Predictor" is where LLMs are responsible for predicting the final answer. "LLM as Aligner" will align the inputs-output pairs with those of GNNs. "LLM as Encoder" refers to using LLMs to encode and obtain feature vectors.

https://bit.ly/ACM24-Colab

Python + SPARQL

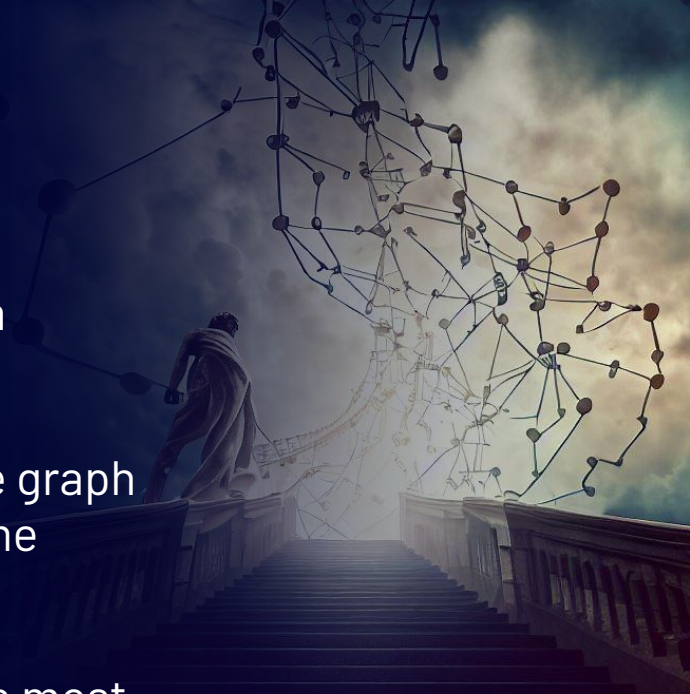
Conclusions

1. The graph model can represent data under different circumstances and different abstraction levels, from simple graph to complex knowledge graphs

2. When to model data as graph? Is the structure of the graph encoding relevant information? What analysis become possible?

3. Proximity, density, connectivity, and patterns are the most common types of information to mine

4. A graph can be an integration layer, or non-graph-data can be transformed into graphs temporarily to extract some insight



References

M. Arenas, G. I. Diaz, and E. V. Kostylev. Reverse engineering sparql queries. WWW, 2016.

Agichtein, E. and Gravano, L. Snowball: Extracting relations from large plain-text collections. ICDL, 2000.

A.Bonifati, R.Ciucanu,and A.Lemay. Learning path queries on graph databases. EDBT, 2015.

A. Bonifati, R. Ciucanu, and S. Staworko. Learning join queries from user examples. TODS, 2016.

A. Bonifati, U. Comignani, E. Coquery, and R. Thion. Interactive mapping specification with exemplar tuples. SIGMOD, 2017.

I. Bordino, G. De Francisci Morales, I. Weber, and F. Bonchi. From machu picchu to rafting the urubamba river: anticipating information needs via the entity-query graph. WSDM, 2013.

D. Deutch and A. Gilad. Qplain: Query by explanation. ICDE, 2016.

References

G. Diaz, M. Arenas, and M. Benedikt. Sparqlbye: Querying rdf data by example. PVLDB, 2016.

K. Dimitriadou, O. Papaemmanouil, and Y. Diao. Explore-by-example: An automatic query steering framework for interactive data exploration. In SIGMOD, 2014.

B. Eravci and H. Ferhatosmanoglu. Diversity based relevance feedback for time series search. PVLDB, 2013.

A. Gionis, M. Mathioudakis, and A. Ukkonen. Bump hunting in the dark: Local discrepancy maximization on graphs. ICDE, 2015.

M. F. Hanafi, A. Abouzied, L. Chiticariu, and Y. Li. Synthesizing extraction rules from user examples with seer. SIGMOD, 2017.

He, J., Veltri, E., Santoro, D., Li, G., Mecca, G., Papotti, P. and Tang, N. Interactive and deterministic data cleaning. SIGMOD, 2016.

Y. Ishikawa, R. Subramanya, and C. Faloutsos. Mindreader: Querying databases through multiple examples. VLDB, 1998.

References

N. Jayaram, A. Khan, C. Li, X. Yan, and R. Elmasri. Querying knowledge graphs by example entity tuples. TKDE, 2015.

H. Li, C.-Y. Chan, and D. Maier. Query from examples: An iterative, data-driven approach to query construction. PVLDB, 2015.

M. Lissandrini, D. Mottin, Y. Velegrakis, T. Palpanas. Multi-Example Search in Rich Information Graphs ICDE 2018

Y. Ma, T.-K. Huang, and J. G. Schneider. Active search and bandits on graphs using sigma-optimality. UAI, 2015.

S. Metzger, R. Schenkel, and M. Sydow. Qbees: query by entity examples. CIKM, 2013.

D. Mottin, M. Lissandrini, Y. Velegrakis, and T. Palpanas. Searching with xq: the exemplar query search engine. SIGMOD, 2014.

D. Mottin, M. Lissandrini, Y. Velegrakis, and T. Palpanas. Exemplar queries: a new way of searching. VLDB J., 2016.

B. Perozzi, L. Akoglu, P. Iglesias Sanchez, and E. Müller. Focused clustering and outlier detection in large attributed graphs. KDD, 2014.

References

F. Psallidas, B. Ding, K. Chakrabarti, and S. Chaudhuri. S4: Top-k spreadsheet-style search for query discovery. SIGMOD, 2015.

R. Rolim, G. Soares, L. D'Antoni, O. Polozov, S. Gulwani, R. Gheyi, R. Suzuki, and B. Hartmann. Learning syntactic program transformations from examples. ICSE, 2017.

N. Ruchansky, F. Bonchi, D. García-Soriano, F. Gullo, and N. Kourtellis. The minimum wiener connector problem. SIGMOD, 2015.

T. Sellam and M. Kersten. Cluster-driven navigation of the query space. TKDE, 2016.

Y. Shen, K. Chakrabarti, S. Chaudhuri, B. Ding, and L. Novik. Discovering queries based on example tuples. SIGMOD, 2014.

R. Singh. Blinkfill: Semi-supervised programming by example for syntactic string transformations. PVLDB, 2016.

G. Sobczak, M. Chochół, R. Schenkel, and M. Sydow. iqbees: Towards interactive semantic entity search based on maximal aspects. Foundations of Intelligent Systems, 2015.

References

Y. Su, S. Yang, H. Sun, M. Srivatsa, S. Kase, M. Vanni, and X. Yan. Exploiting relevance feedback in knowledge graph search. KDD, 2015.

Q. T. Tran, C.-Y. Chan, and S. Parthasarathy. Query reverse engineering. VLDB J., 2014.

H. P. Vanchinathan, A. Marfurt, C.-A. Robelin, D. Koss- mann, and A. Krause. Discovering valuable items from massive data. In KDD, 2015.

C.Wang, A.Cheung, and R.Bodik. Interactive query synthesis from input-output examples. In SIGMOD, 2017.

C. Wang, A. Cheung, and R. Bodik. Synthesizing highly expressive sql queries from input-output examples. In PLDI, 2017.

Y. Y. Weiss and S. Cohen. Reverse engineering spj-queries from examples. SIGMOD, 2017.

M. Yakout, K. Ganjam, K. Chakrabarti, and S. Chaudhuri. Infogather: Entity augmentation and attribute discovery by holistic matching with web tables. SIGMOD, 2012.

M. Zhu and Y.-F. B. Wu. Search by multiple examples. WSDM, 2014.

M. M. Zloof. Query by example. AFIPS NCC, 1975.