THE KNOWLEDGE GRAPH JOURNEY, FROM DATA TO REASONING

Romain Fontugne - iijlab seminar – 2024/10/29

AGENDA

Why knowledge graphs?

Popular knowledge graphs

Example applications

IYP: Knowledge graph for the Internet

DATA, DATA, DATA (AND METADATA)

- **Data collection everywhere!**
- **From different sources: application-driven, monitoring, tracking, survey, experiments**
- **In different forms:**
  Database, cloud, data store / lake / silos / warehouse

→ Data is valuable, make good use of yours

The Economist | Menu | Try for free | Log in

Leaders | Regulating the internet giants

The world's most valuable resource is no longer oil, but data

The data economy demands a new approach to antitrust rules

May 6th 2017 Share

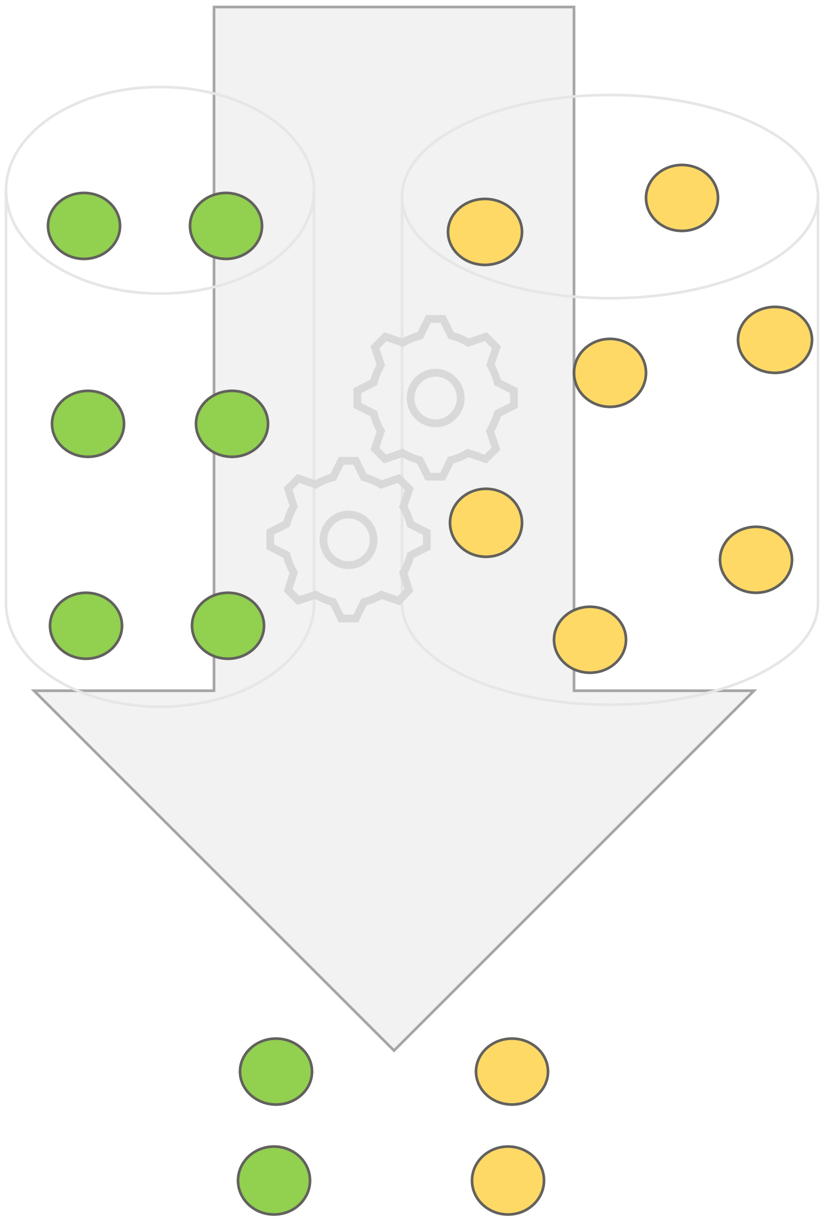
A NEW commodity spawns a lucrative, fast-growing industry, prompting antitrust regulators to step in to restrain those who control its flow. A century ago, the resource in question was oil. Now similar concerns are being raised by the giants that deal in data, the oil of the digital era. These titans—Alphabet (Google's parent company), Amazon, Apple, Facebook and Microsoft—look unstoppable. They are the five most valuable listed firms in the world. Their profits are surging: they collectively racked up over $25bn in net profit in the first quarter of 2017. Amazon captures half of all dollars spent online in America. Google and Facebook accounted for almost all the revenue growth in digital advertising in America last year.

CROSS ANALYSIS

How to get more from your data?

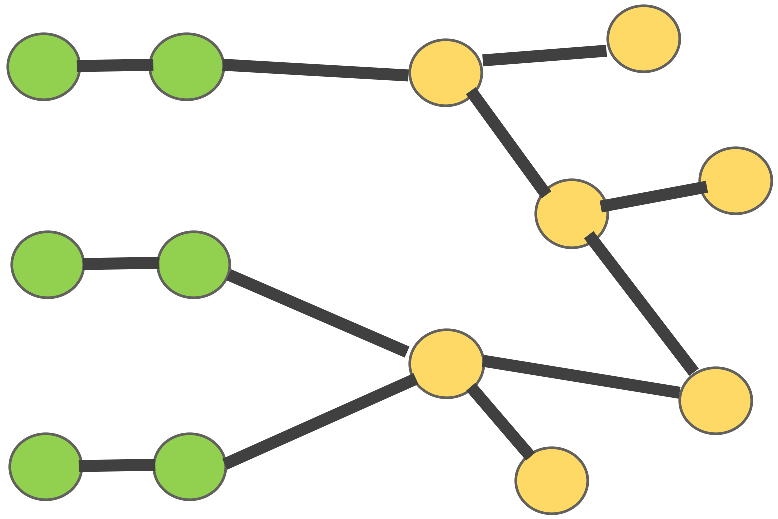
- **Implement data pipelines ingesting different datasets**
- **Custom-made:**
  - **Usually serving a single purpose**
  - **Built for efficiency**

→ Usually efficient but not flexible



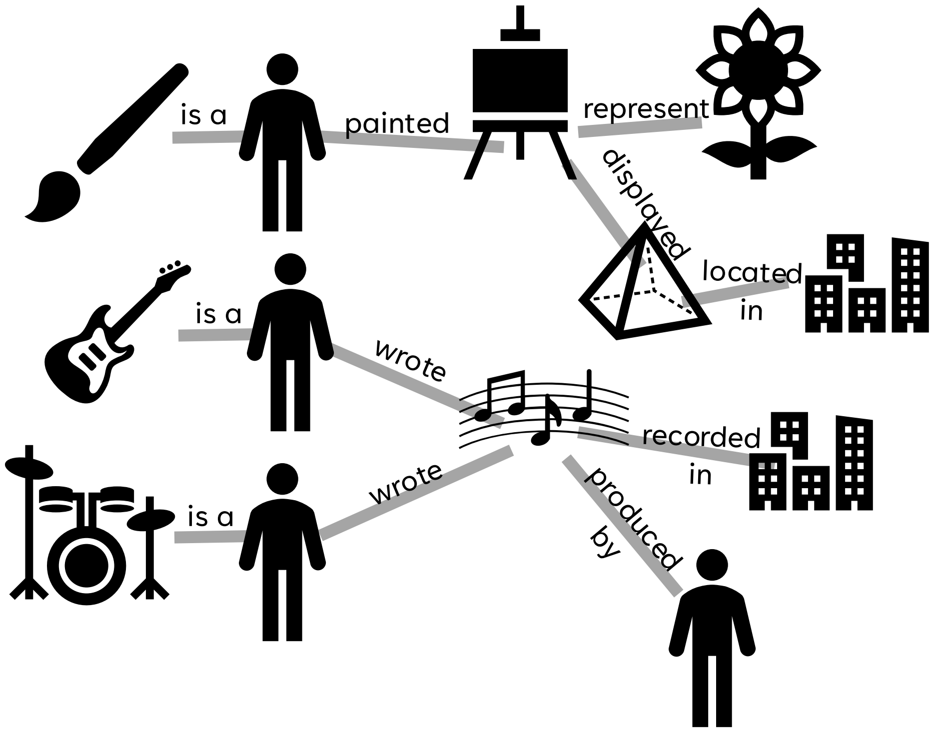
KNOWLEDGE GRAPH

- **Generalization of cross analysis**
  - Graph-structured data model
  - Semantics (ontology):
    - Nodes / Entities
    - Edges / Relationships



KNOWLEDGE GRAPH

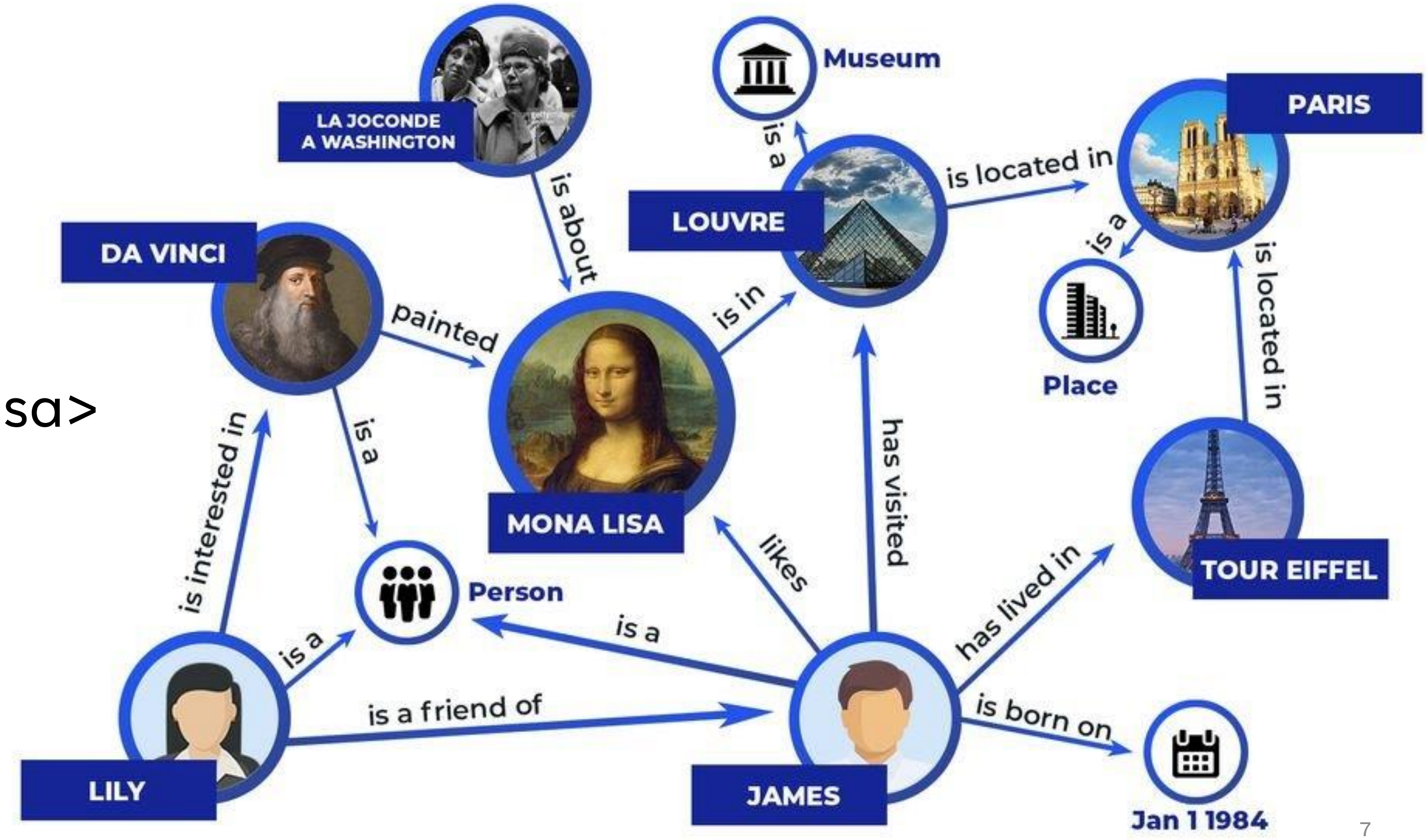
- **Generalization of cross analysis**
  - Graph-structured data model
  - Semantics (ontology):
    - Nodes / Entities
    - Edges / Relationships



→ Self-explainable data structure!

KNOWLEDGE GRAPH: DEFINITIONS

- **From Wikipedia:** *"There is no single commonly accepted definition of a knowledge graph."*
- **Entities:** objects, events, or abstract concepts
- **Relationships**
- **Facts (triples):**
  <Da Vinci, painted, Mona Lisa>



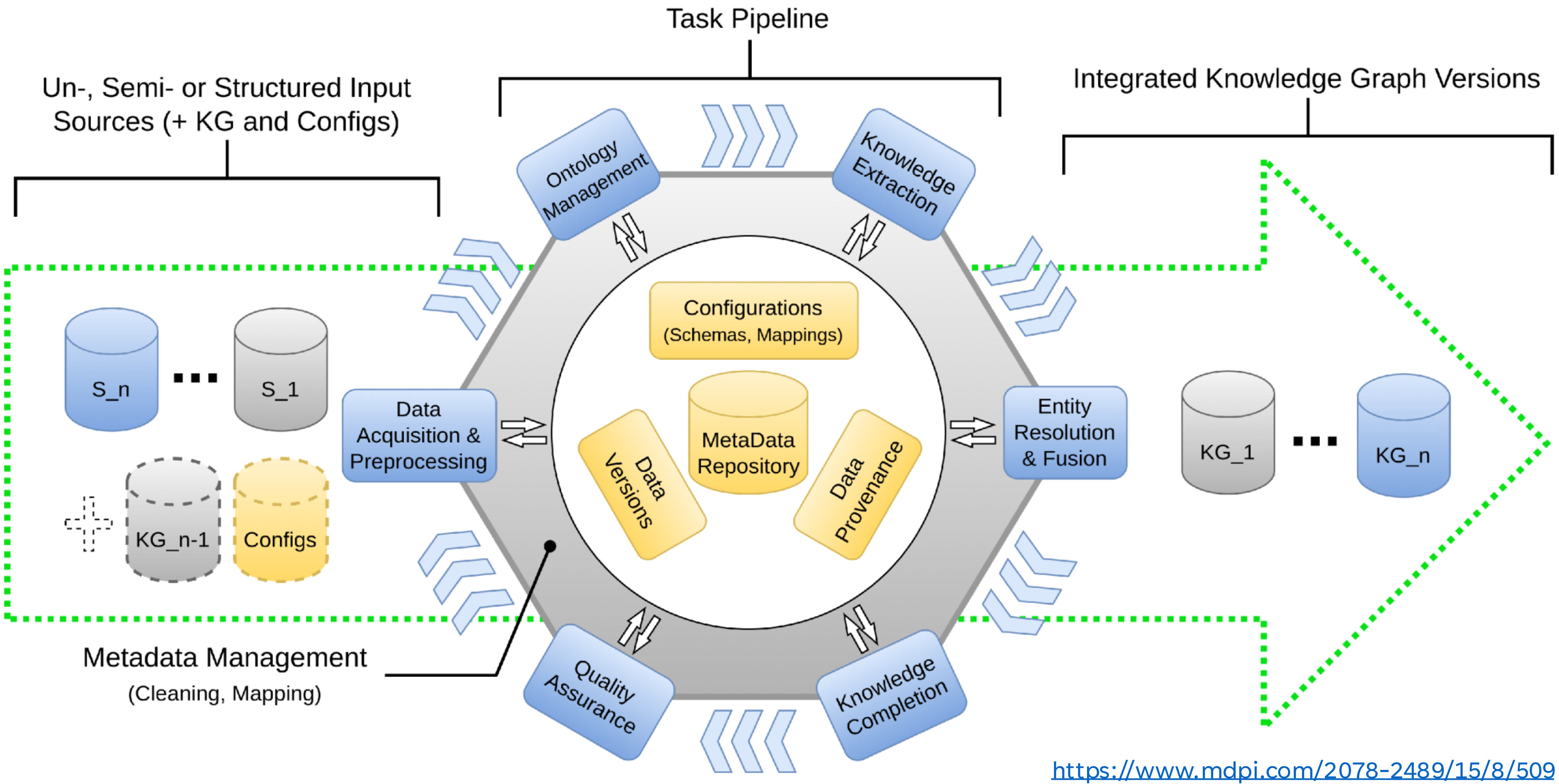
CREATING A KNOWLEDGE GRAPH

- **Ontology:** Naming, hierarchy, level of detail

  "there is no single correct ontology for any domain"

- **Knowledge Acquisition**
  - Entity/Relation extraction
  - Attribute extraction

- **Knowledge Fusion**
  - Entity Alignment: John Smith = J. Smith?
  - Disagreeing datasets?



https://www.mdpi.com/2078-2489/15/8/509

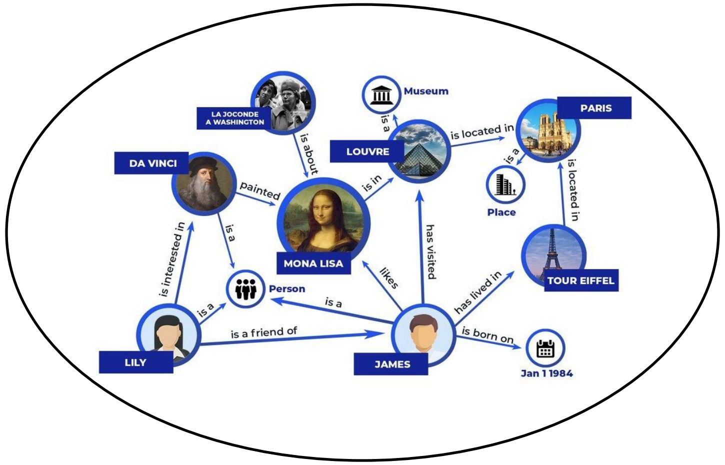
EXAMPLE KNOWLEDGE GRAPH

	Entities	Facts	Comments
Google (2022)	5 billion	500 billion	Google search
Microsoft (2019)	2 billion	55 billion	Bing, Academic, LinkedIn
Facebook (2019)	50 million	500 million	Rebuilt every day
Wikidata (2023)	100 million	15 billion	+10k relationship types manually curated

And a lot more: Netflix, Amazon, eBay, IBM, NASA, …

KNOWLEDGE GRAPH COMMON USES

- **Information retrieval**
- **Semantic search**



- **Question answering systems**

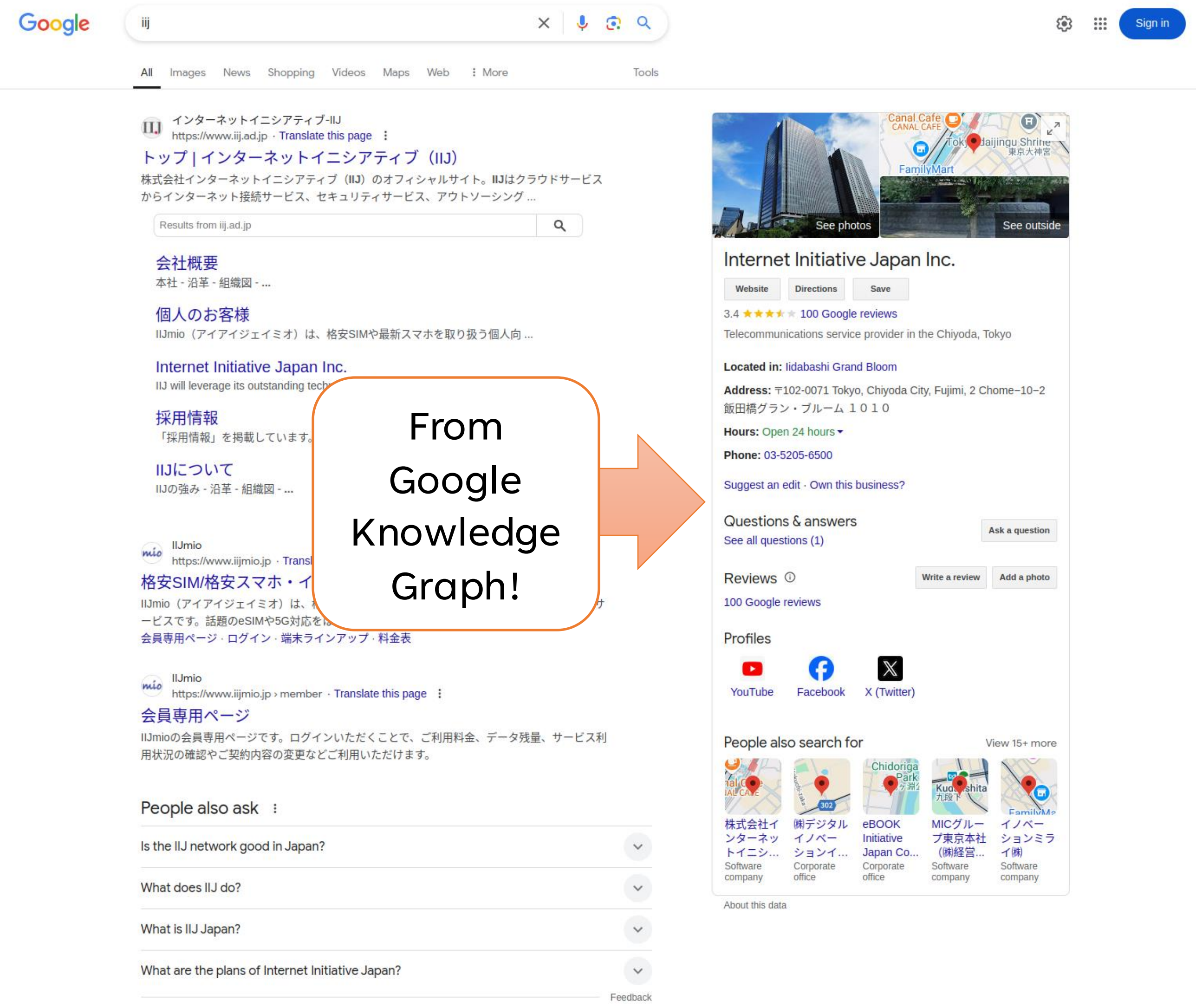
- **Reasoning**
  - Graph traversal
  - Embedding (recommendation)

- **Domain specific**

INFORMATION RETRIEVAL

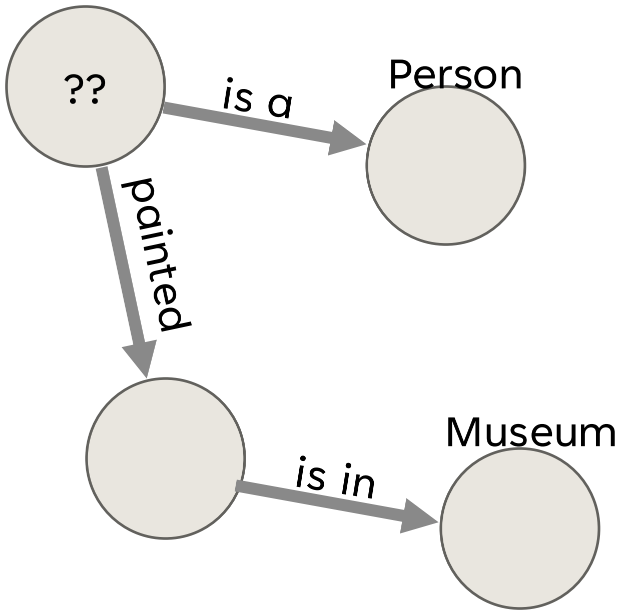
- **Example: Google Search**
- *"things, not strings"*
  - Find the right thing
  - Get the best summary
  - Go deeper and broader

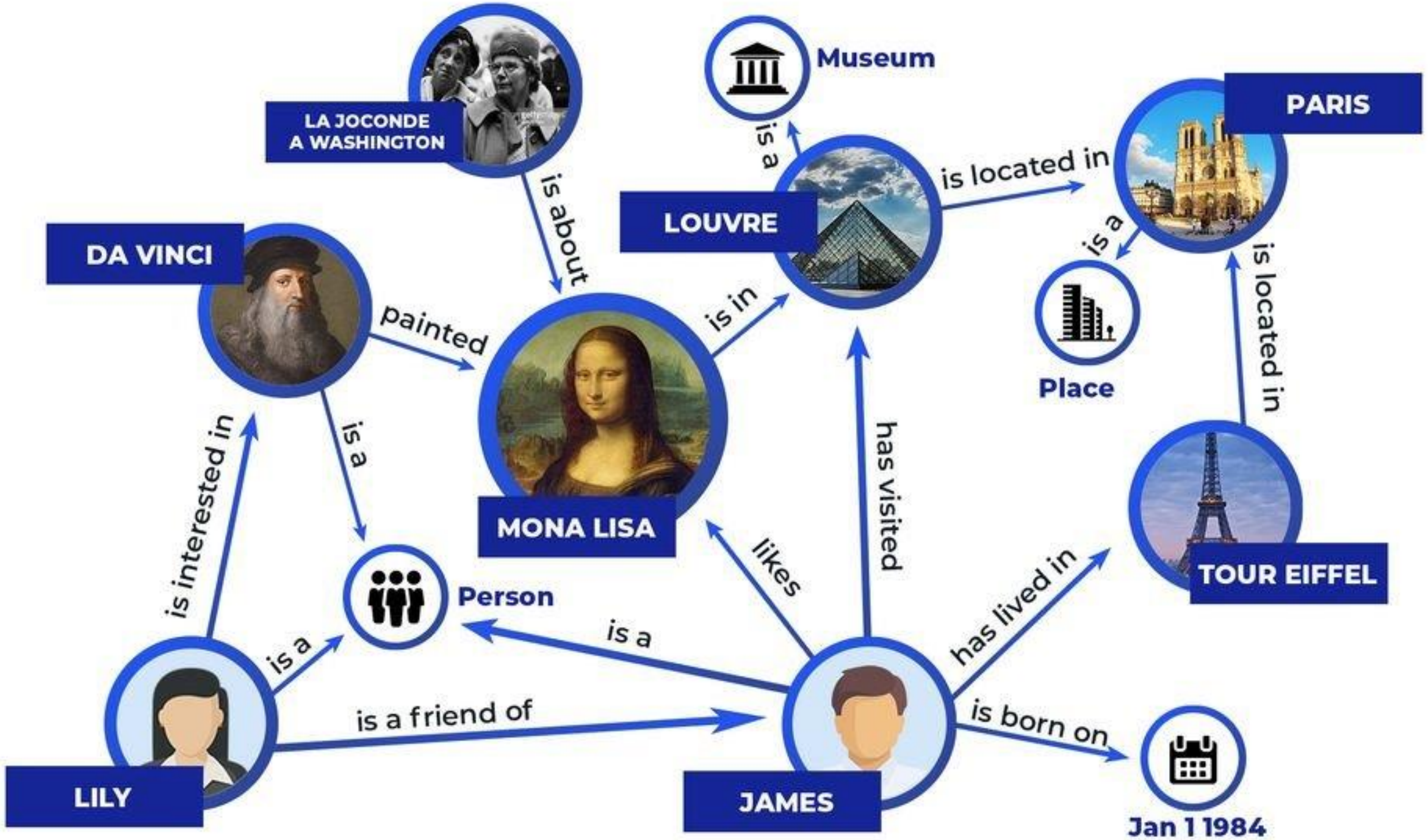
Also allow exploratory search



SEMANTIC SEARCH

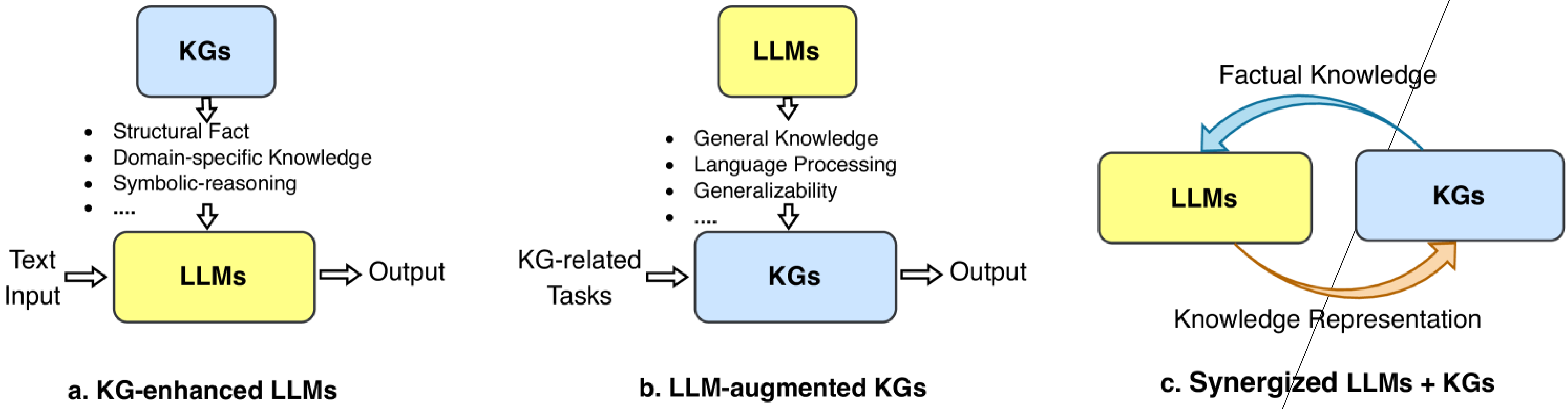
- **Looking for semantic patterns, not keywords!**
- **Include context & user intent**





QUESTION ANSWERING

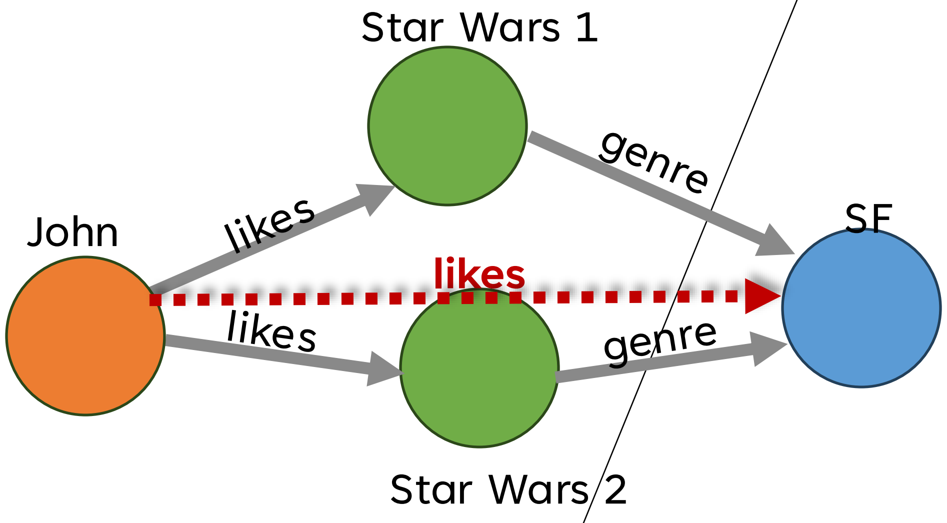
- **Example: IBM Watson, Siri, OK Google**
- **Graph RAG: Help LLMs by giving them context**



Source: https://arxiv.org/html/2306.08302v3

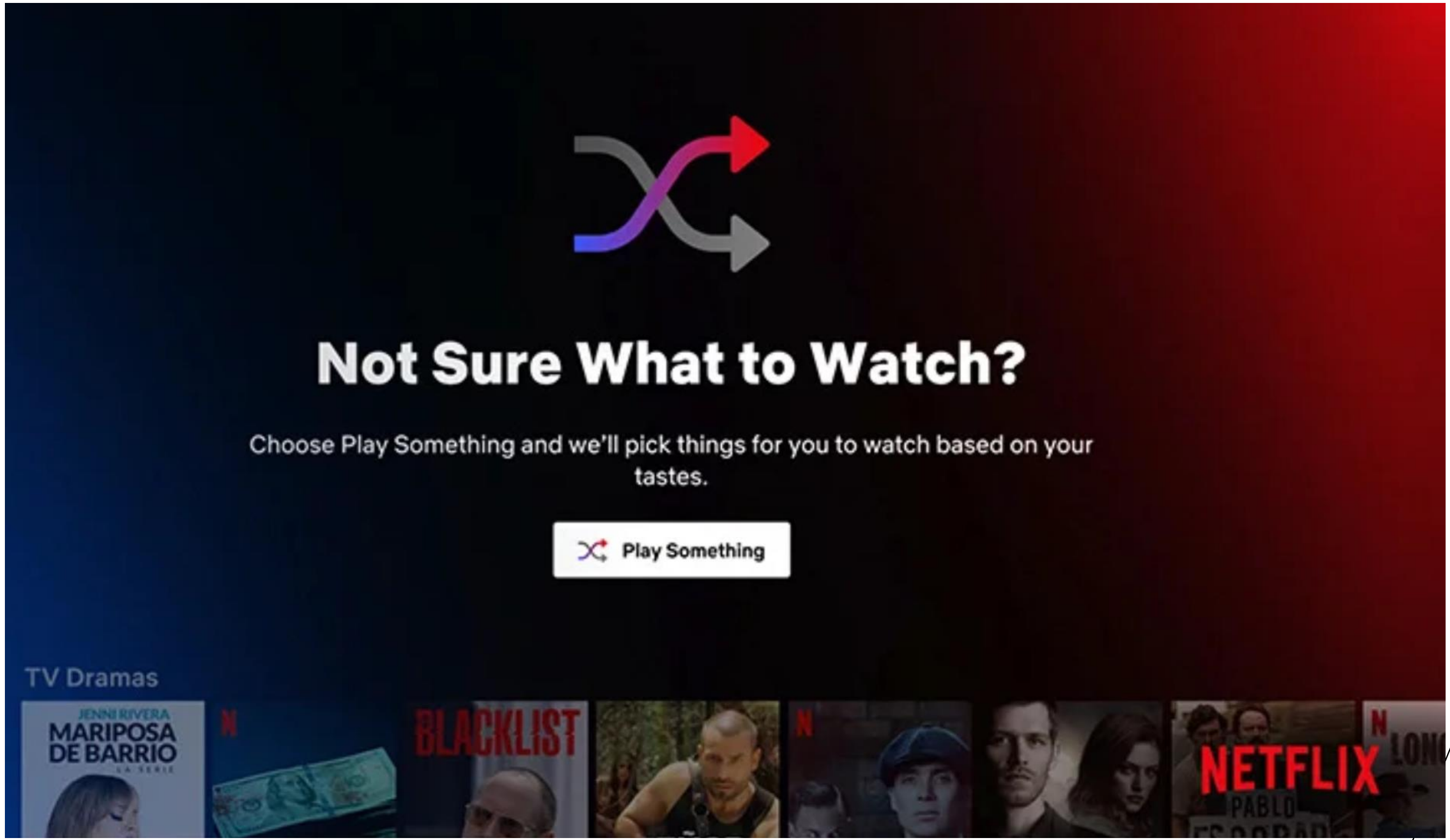
REASONING: GRAPH TRAVERSAL

- **Based on explicit knowledge**
  - Deduce new knowledge
  - Identify wrong knowledge



REASONING: EMBEDDING

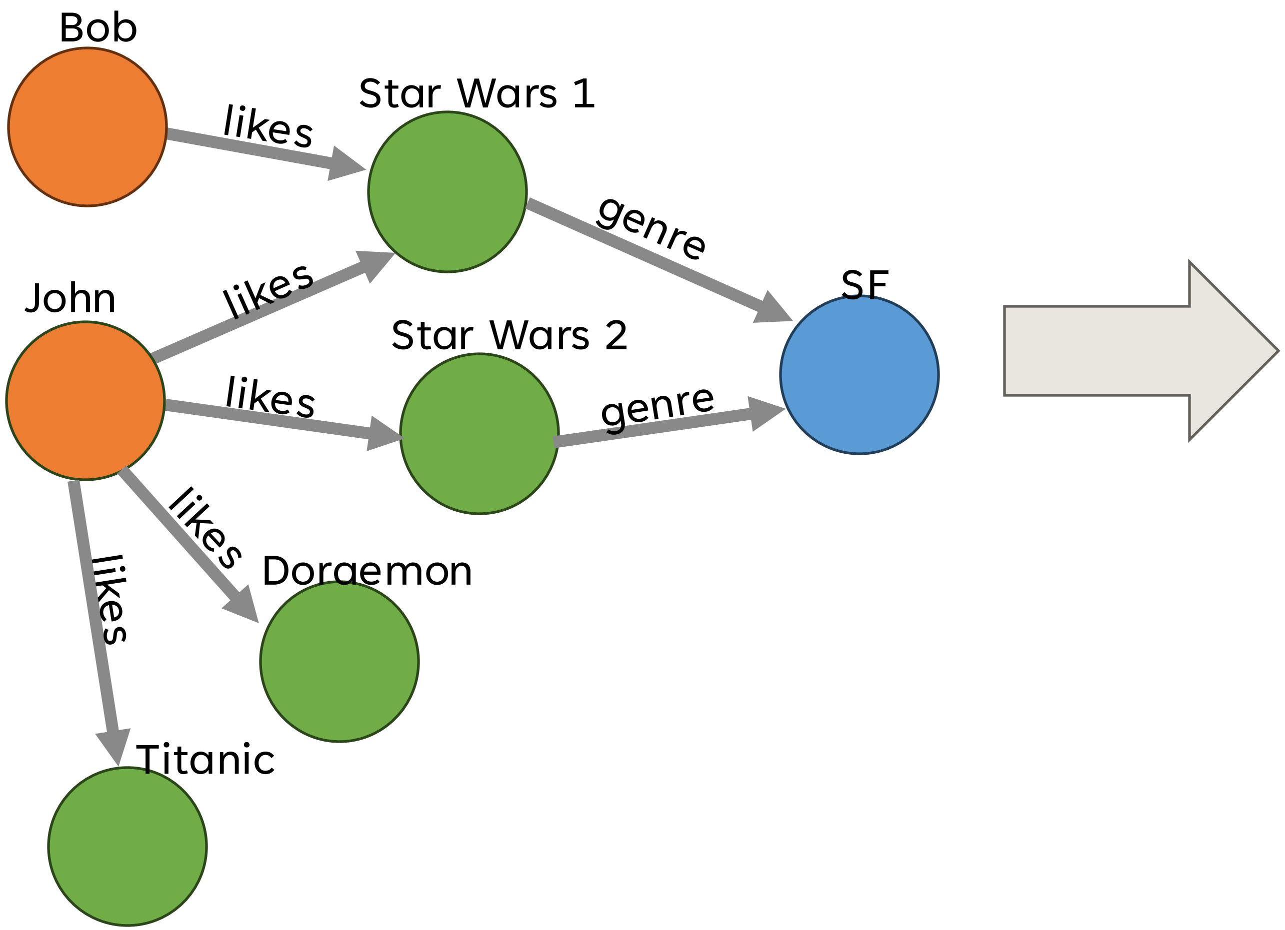
- **Example: Recommendation systems**
  Netflix, Amazon, Facebook
- **The graph is too constrained, project data in a latent space**

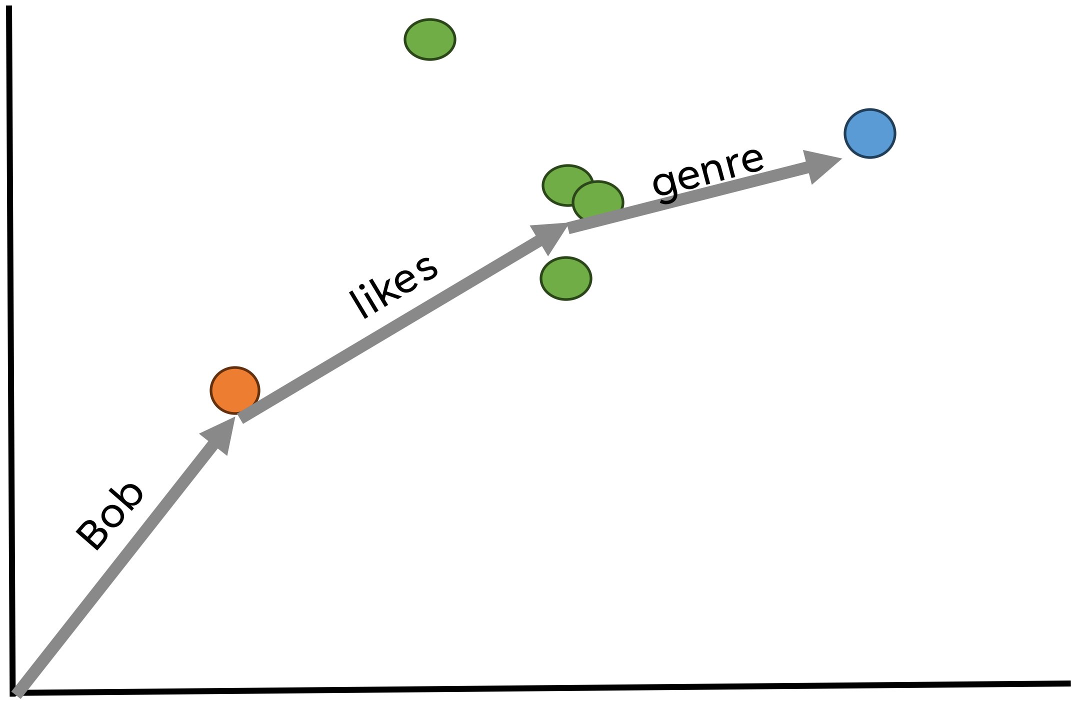


EMBEDDING: OVERVIEW

Input Knowledge Graph

Output KG Embeddings

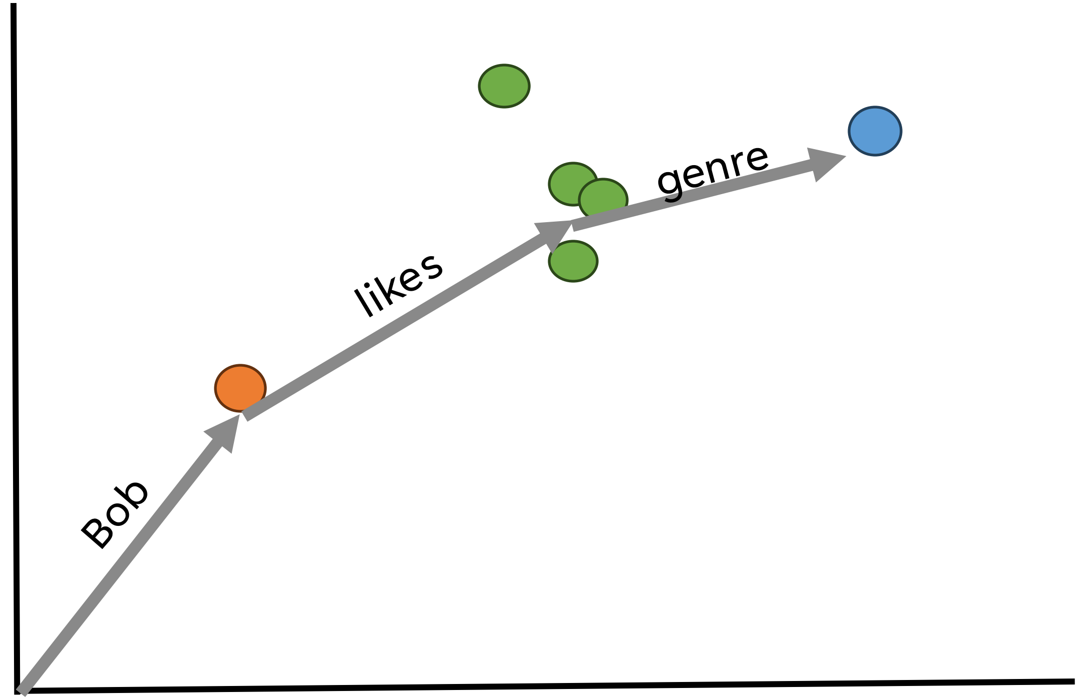




EMBEDDING: APPLICATIONS

Helpful for numerous tasks:

- **Similarity**
- **Clustering**
- **Classification**
- **Link prediction / recommendation**



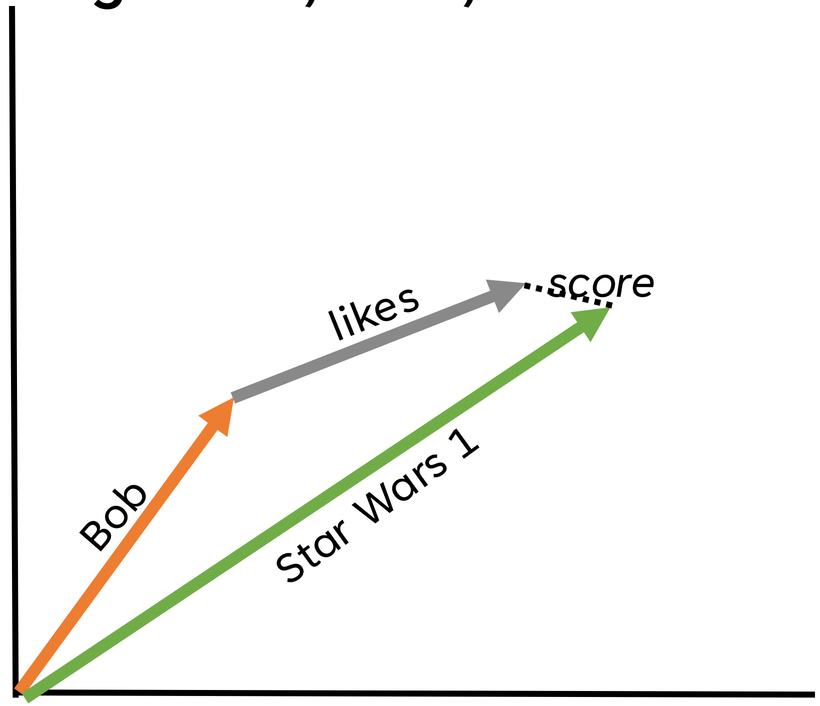
EMBEDDING: ALGORITHMS

How to find good embeddings?

- **Learning from facts: <h, r, t>**

Model	Scoring Function
TransE	$\|\boldsymbol{h}+\boldsymbol{r}-\boldsymbol{t}\|_i$
TransH	$\|(\boldsymbol{h}-\boldsymbol{w}_r^{\mathrm{T}}\boldsymbol{h}\boldsymbol{w}_r)+\boldsymbol{r}-(\boldsymbol{t}-\boldsymbol{w}_r^{\mathrm{T}}\boldsymbol{t}\boldsymbol{w}_r)\|_i$
TransR	$\|\boldsymbol{W}_r\boldsymbol{h}+\boldsymbol{r}-\boldsymbol{W}_r\boldsymbol{t}\|_i$
RESCAL	$\boldsymbol{h}^{\mathrm{T}}\boldsymbol{W}_r\boldsymbol{t}$
DistMult	$\boldsymbol{h}^{\mathrm{T}}diag(\boldsymbol{r})\boldsymbol{t}$
ComplEx	$Re(\boldsymbol{h}^{\mathrm{T}}diag(\boldsymbol{r})\overline{\boldsymbol{t}})$

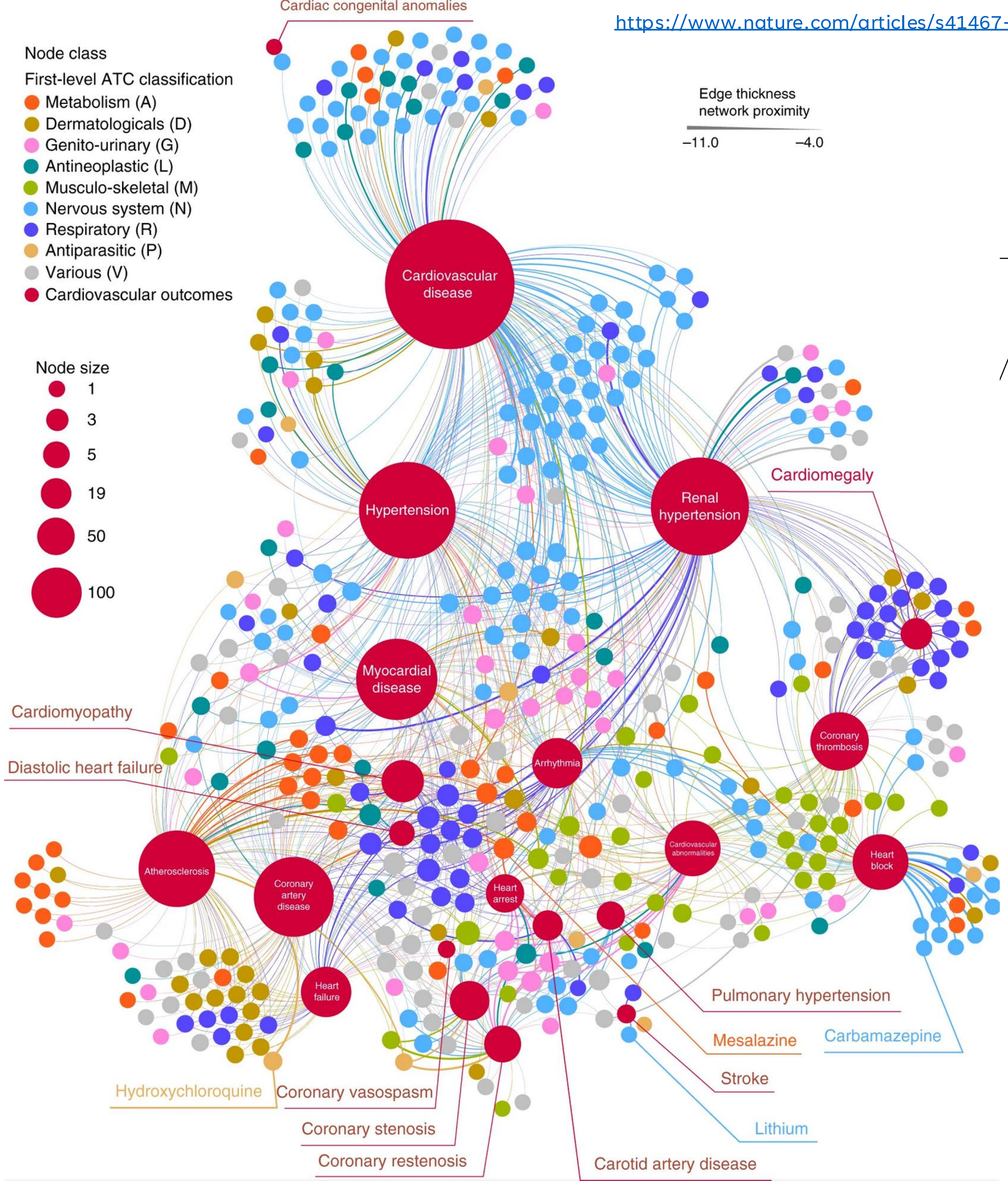
e.g. <Bob, likes, Star Wars 1>



DOMAIN SPECIFIC

- **e.g. Medical applications**
  - DrugBank
  - PharmGKB
  - Gene ontology

https://www.nature.com/articles/s41467-018-05116-5

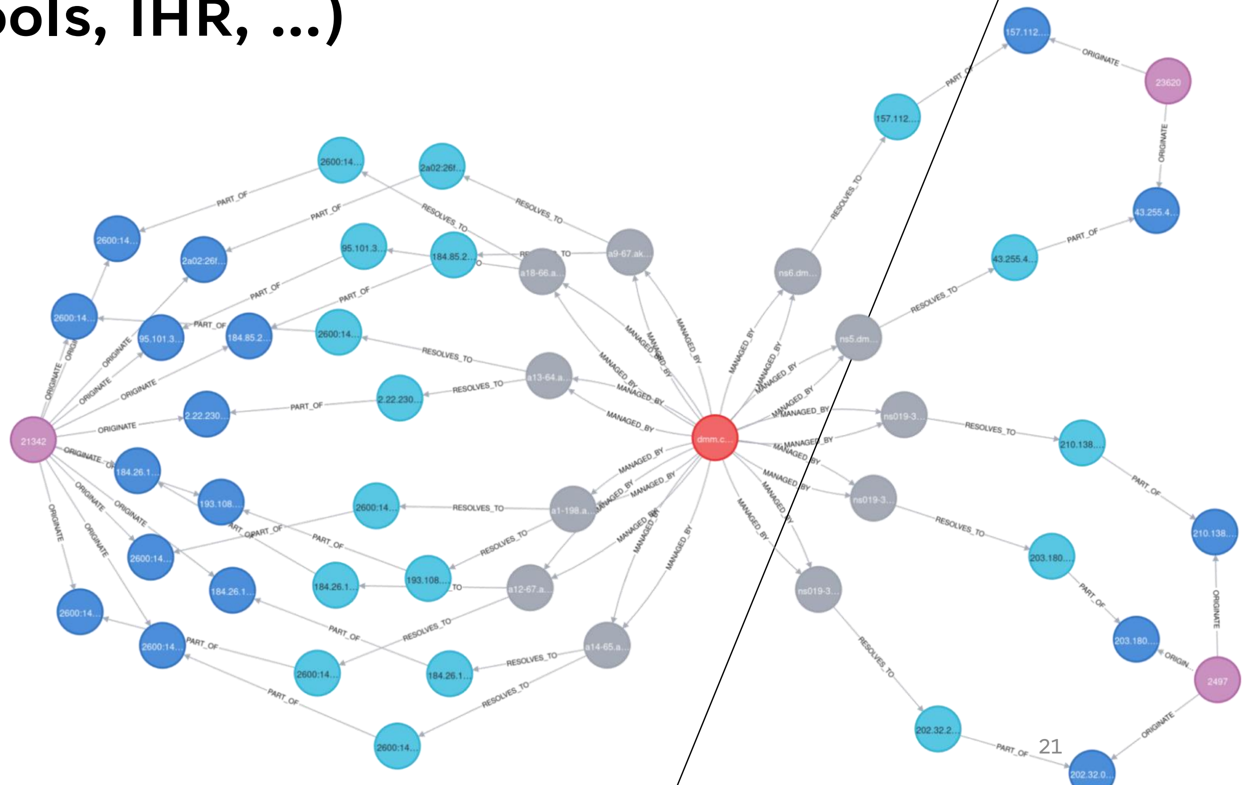


INTERNET YELLOW PAGES

Knowledge graph for Internet:

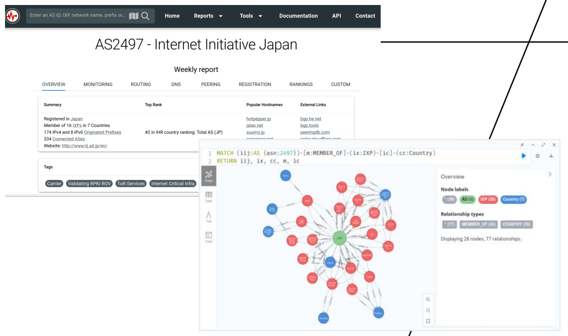
- **40+ datasets (PeeringDB, CAIDA, RIPE, APNIC, Cloudflare, OONI, BGPKit, BGP.Tools, IHR, …)**
- **12+ million nodes**
- **82+ million edges**
- **Produced every week**
- **Available online at:**

  https://iyp.iijlab.net

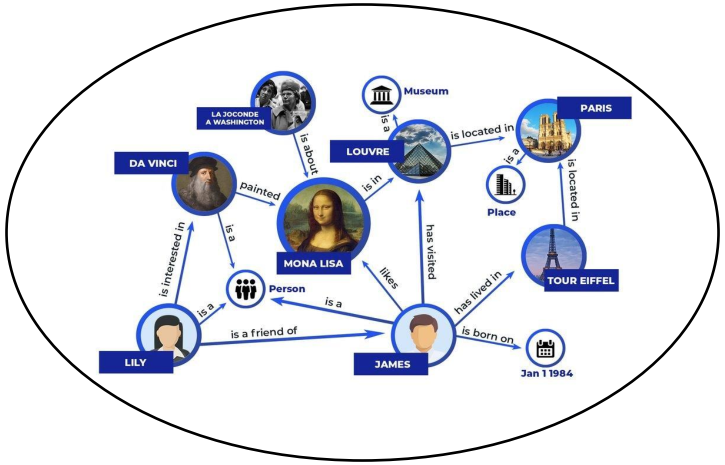


IYP APPLICATIONS

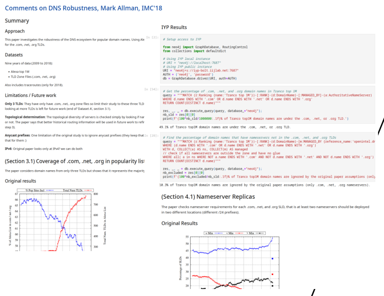
- **Information retrieval**
- **Semantic search**



- **Question answering systems**

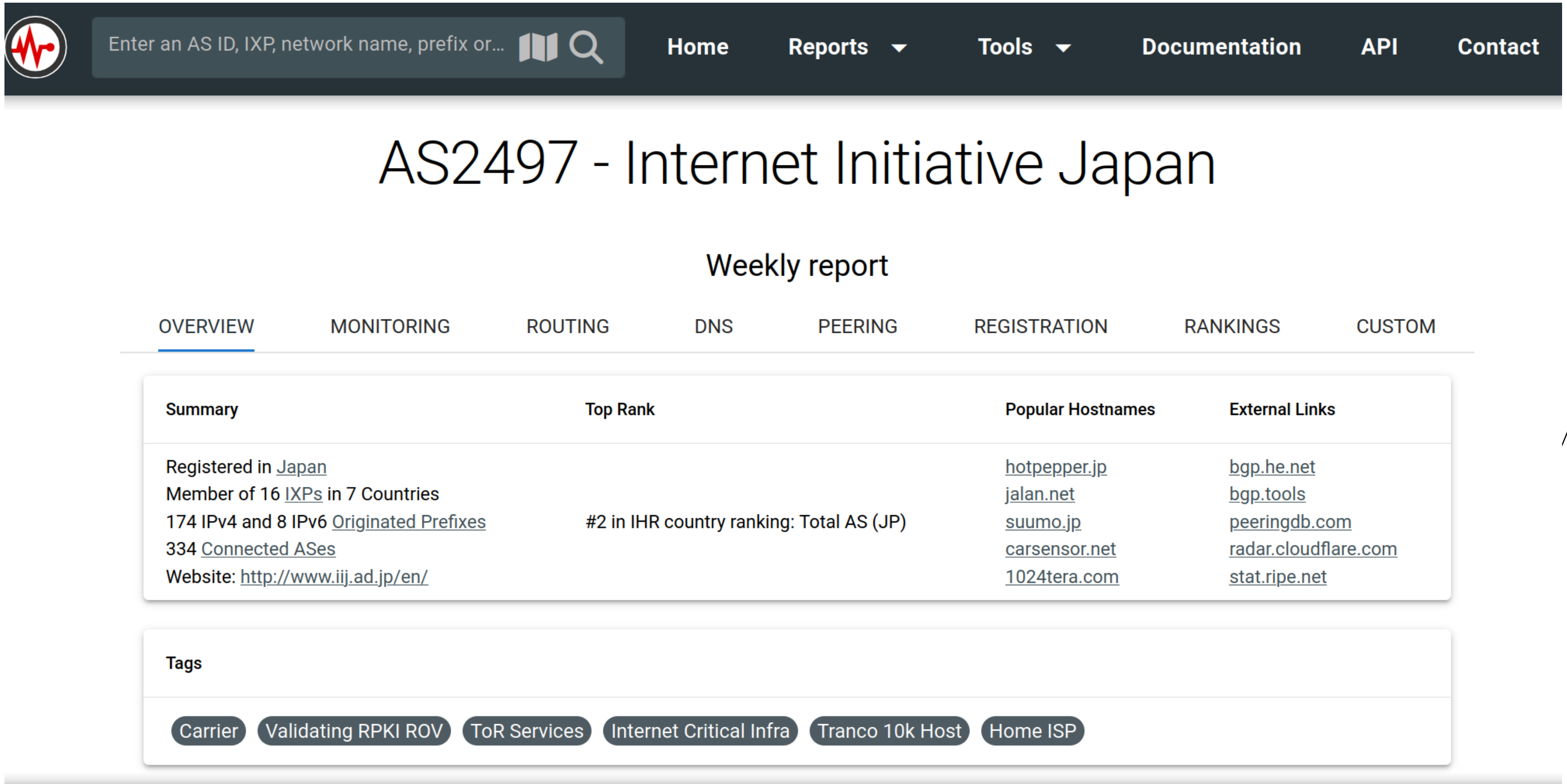


- **Reasoning**
  - Graph traversal
  - Embedding



IYP: INFORMATION RETRIEVAL

- https://ihr.iijlab.net
  - Simple search
  - Predefined templates

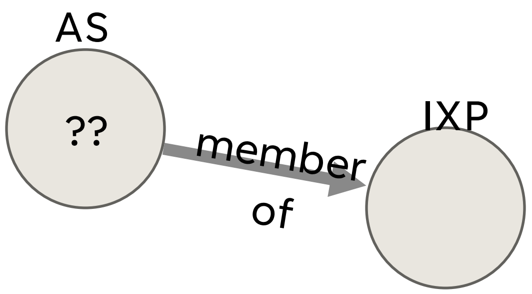


SEMANTIC SEARCH

https://iyp.iijlab.net

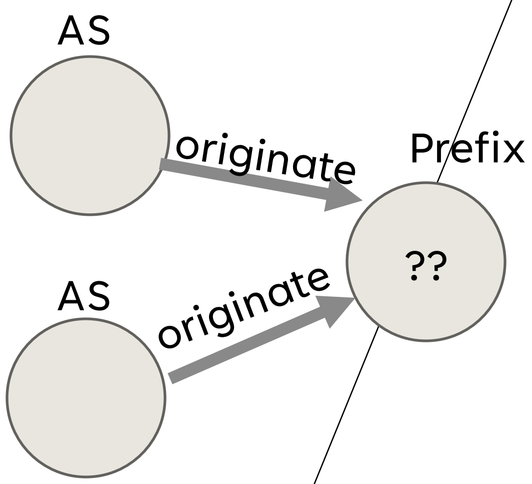
- **AS member of an IXP?**

```
MATCH (a:AS)-[:MEMBER_OF]-(ix:IXP)
RETURN COUNT(DISTINCT a)
```



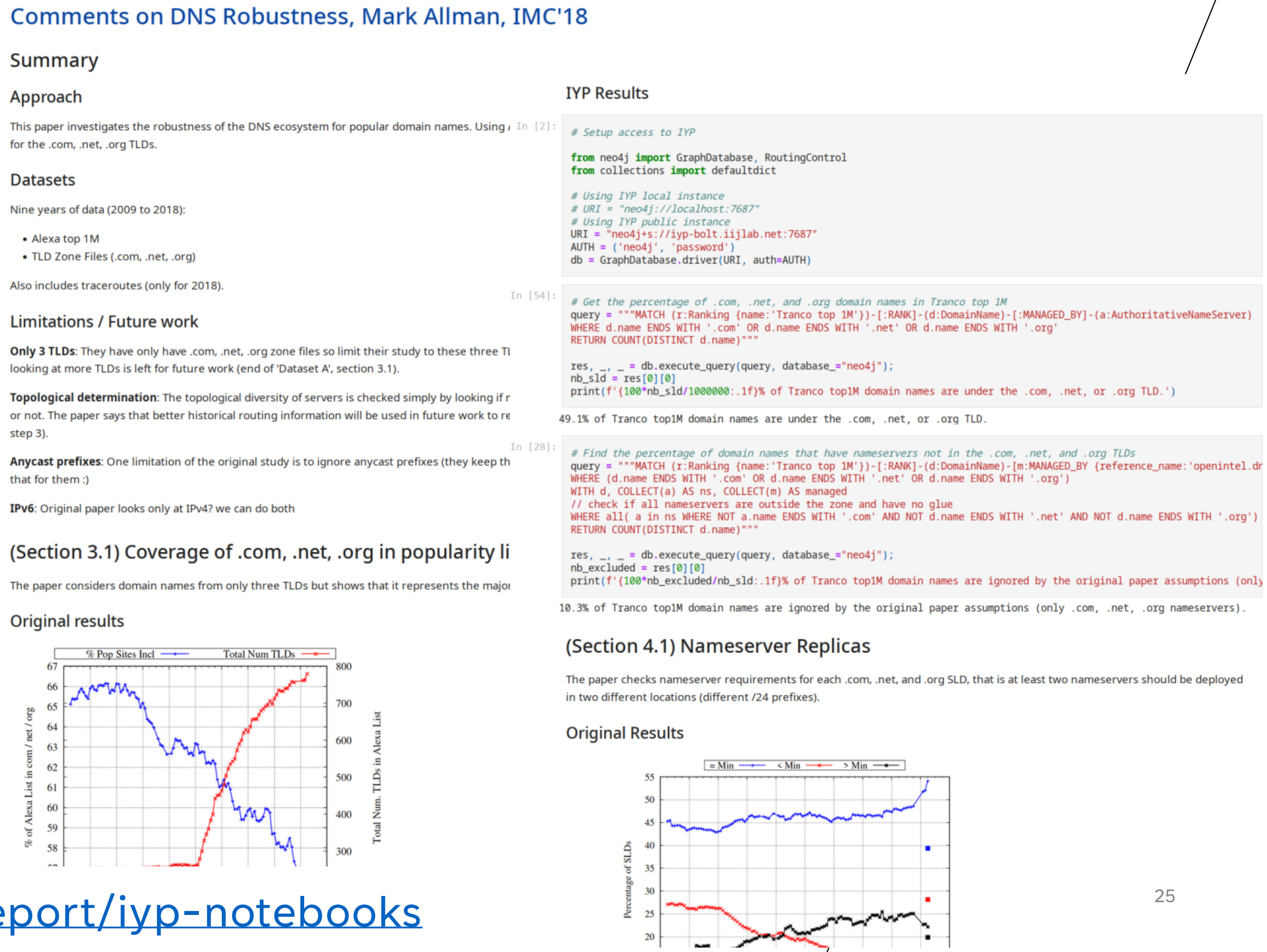
- **Prefixes originated by multiple ASes**

```
MATCH (a:AS)-[:ORIGINATE]-(p:Prefix)-[:ORIGINATE]-(b:AS)
WHERE a <> b
RETURN COUNT(DISTINCT p)
```



IYP: REASONING (GRAPH TRAVERSAL)

- **DNS robustness (IMC'18)**
- **RPKI deployment (HotNets'15)**



Comments on DNS Robustness, Mark Allman, IMC'18

Summary

Approach

This paper investigates the robustness of the DNS ecosystem for popular domain names. Using a
for the .com, .net, .org TLDs.

Datasets

Nine years of data (2009 to 2018):

- Alexa top 1M
- TLD Zone Files (.com, .net, .org)

Also includes traceroutes (only for 2018).

Limitations / Future work

Only 3 TLDs: They have only have .com, .net, .org zone files so limit their study to these three Tl
looking at more TLDs is left for future work (end of 'Dataset A', section 3.1).

Topological determination: The topological diversity of servers is checked simply by looking if r
or not. The paper says that better historical routing information will be used in future work to re
step 3).

Anycast prefixes: One limitation of the original study is to ignore anycast prefixes (they keep th
that for them :)

IPv6: Original paper looks only at IPv4? we can do both

(Section 3.1) Coverage of .com, .net, .org in popularity li

The paper considers domain names from only three TLDs but shows that it represents the majoı

Original results

IYP Results

In [2]:
```python
# Setup access to IYP

from neo4j import GraphDatabase, RoutingControl
from collections import defaultdict

# Using IYP local instance
# URI = "neo4j://localhost:7687"
# Using IYP public instance
URI = "neo4j+s://iyp-bolt.iijlab.net:7687"
AUTH = ('neo4j', 'password')
db = GraphDatabase.driver(URI, auth=AUTH)
```

In [54]:
```python
# Get the percentage of .com, .net, and .org domain names in Tranco top 1M
query = """MATCH (r:Ranking {name:'Tranco top 1M'})-[:RANK]-(d:DomainName)-[:MANAGED_BY]-(a:AuthoritativeNameServer)
WHERE d.name ENDS WITH '.com' OR d.name ENDS WITH '.net' OR d.name ENDS WITH '.org'
RETURN COUNT(DISTINCT d.name)"""

res, _, _ = db.execute_query(query, database_="neo4j");
nb_sld = res[0][0]
print(f'{100*nb_sld/1000000:.1f}% of Tranco top1M domain names are under the .com, .net, or .org TLD.')
```

```
49.1% of Tranco top1M domain names are under the .com, .net, or .org TLD.
```

In [28]:
```python
# Find the percentage of domain names that have nameservers not in the .com, .net, and .org TLDs
query = """MATCH (r:Ranking {name:'Tranco top 1M'})-[:RANK]-(d:DomainName)-[m:MANAGED_BY {reference_name:'openintel.dr
WHERE (d.name ENDS WITH '.com' OR d.name ENDS WITH '.net' OR d.name ENDS WITH '.org')
WITH d, COLLECT(a) AS ns, COLLECT(m) AS managed
// check if all nameservers are outside the zone and have no glue
WHERE all( a in ns WHERE NOT a.name ENDS WITH '.com' AND NOT d.name ENDS WITH '.net' AND NOT d.name ENDS WITH '.org')
RETURN COUNT(DISTINCT d.name)"""

res, _, _ = db.execute_query(query, database_="neo4j");
nb_excluded = res[0][0]
print(f'{100*nb_excluded/nb_sld:.1f}% of Tranco top1M domain names are ignored by the original paper assumptions (only
```

```
10.3% of Tranco top1M domain names are ignored by the original paper assumptions (only .com, .net, .org nameservers).
```

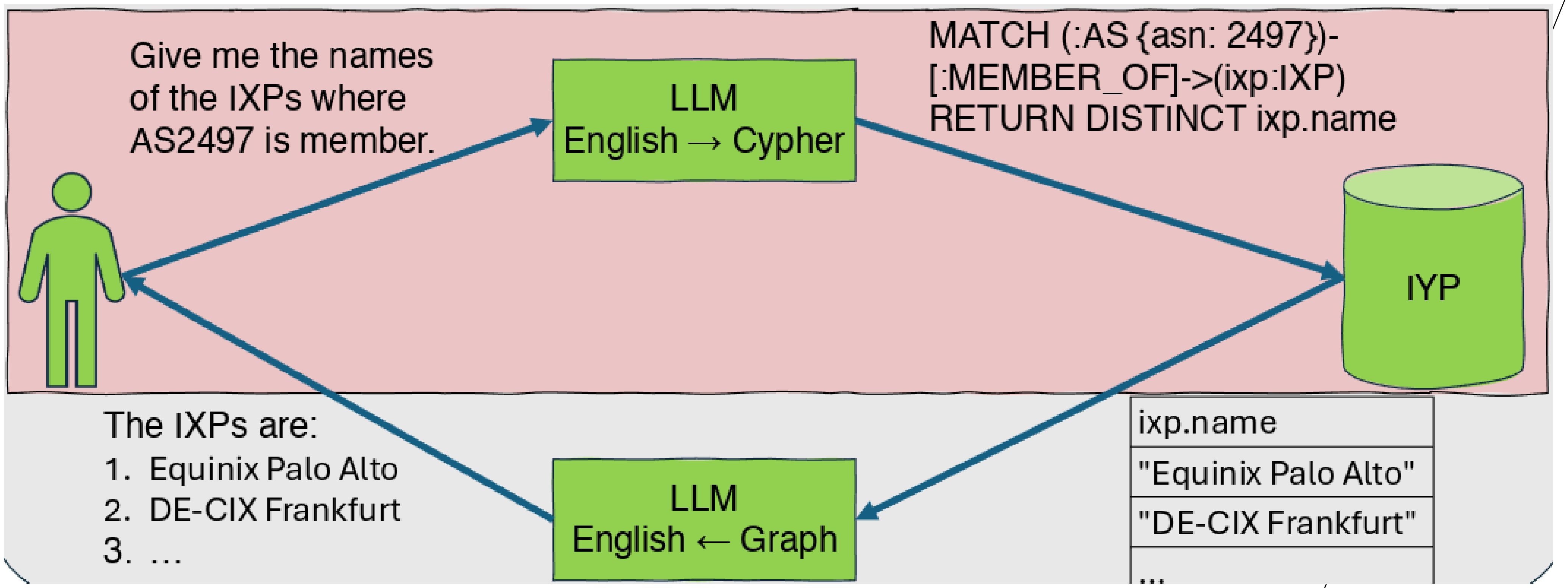
(Section 4.1) Nameserver Replicas

The paper checks nameserver requirements for each .com, .net, and .org SLD, that is at least two nameservers should be deployed in two different locations (different /24 prefixes).

Original Results

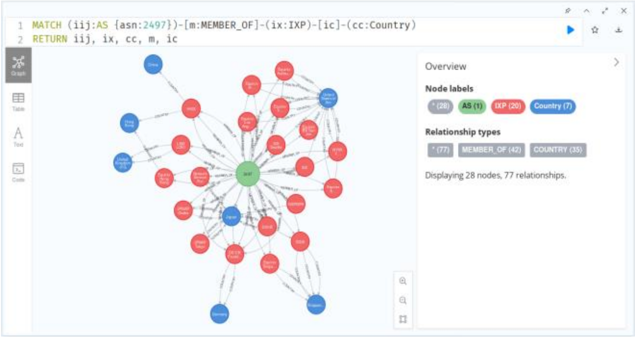
https://github.com/InternetHealthReport/iyp-notebooks

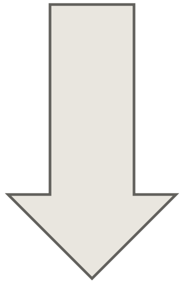
IYP: QUESTION ANSWERING SYSTEM

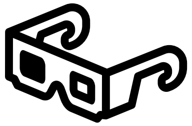


NEXT STEPS

- Recommendation systems
  - Peering recommendations
    - In a specific region?
    - For a specific industry?
  - AS classification
  - Country similarities

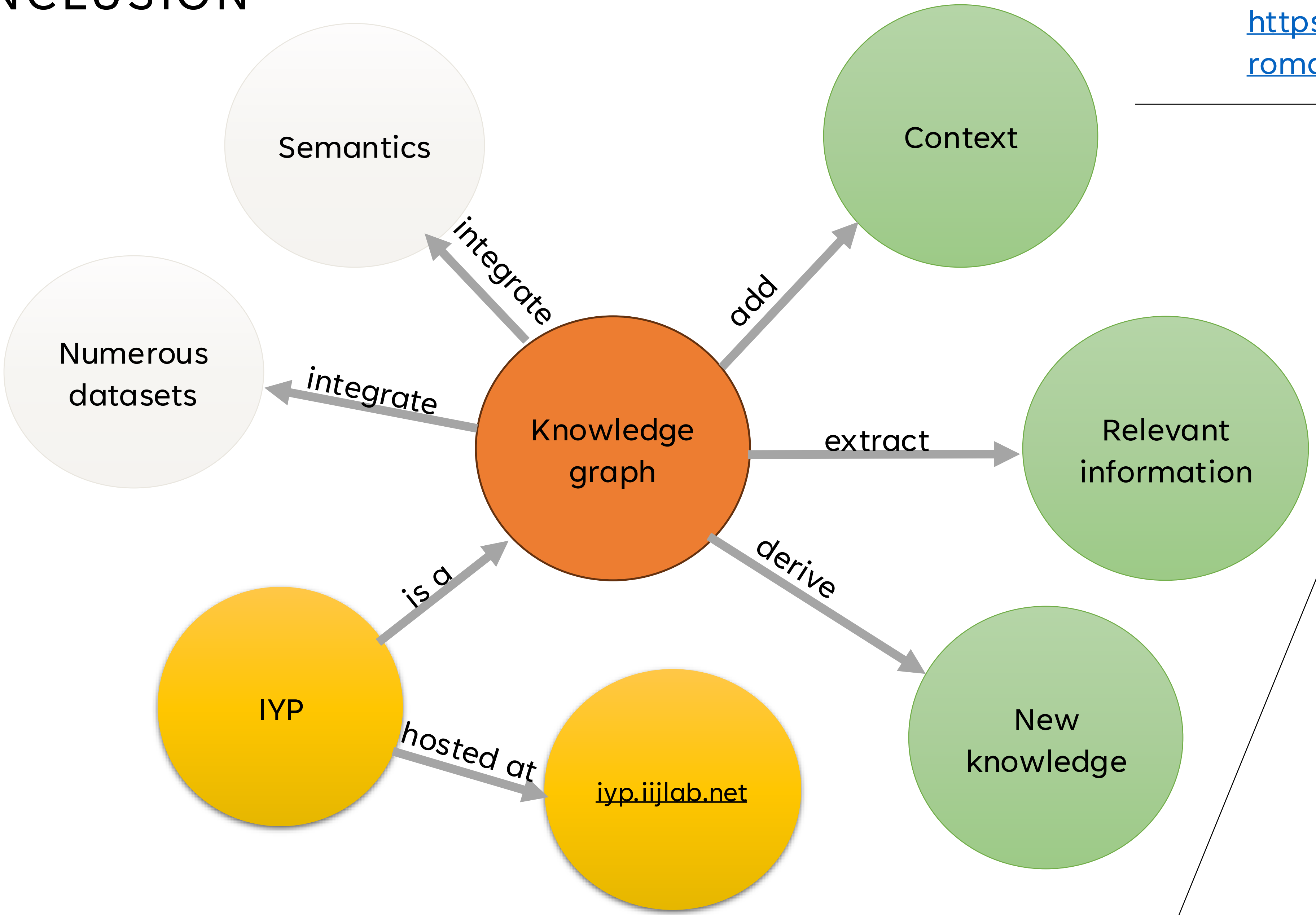






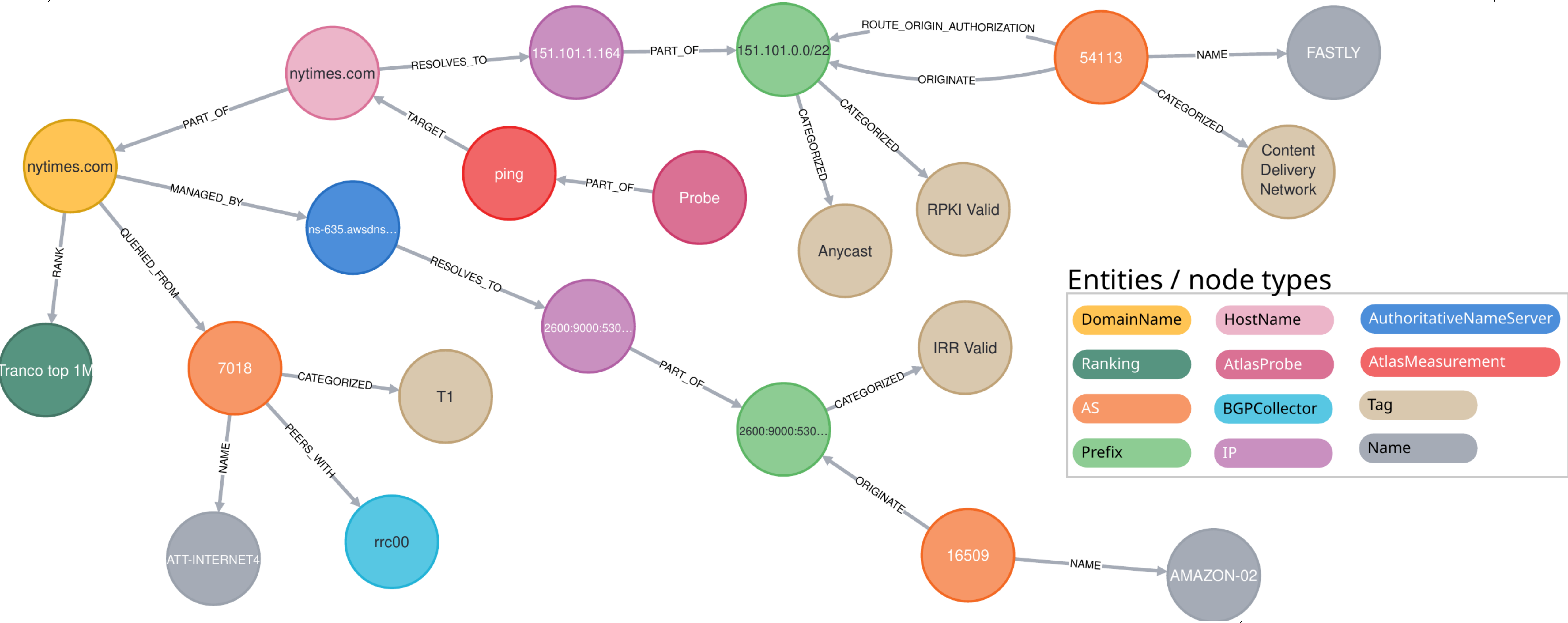
CONCLUSION

https://ihr.iijlab.net
romain@iij.ad.jp



BACKUP SLIDES

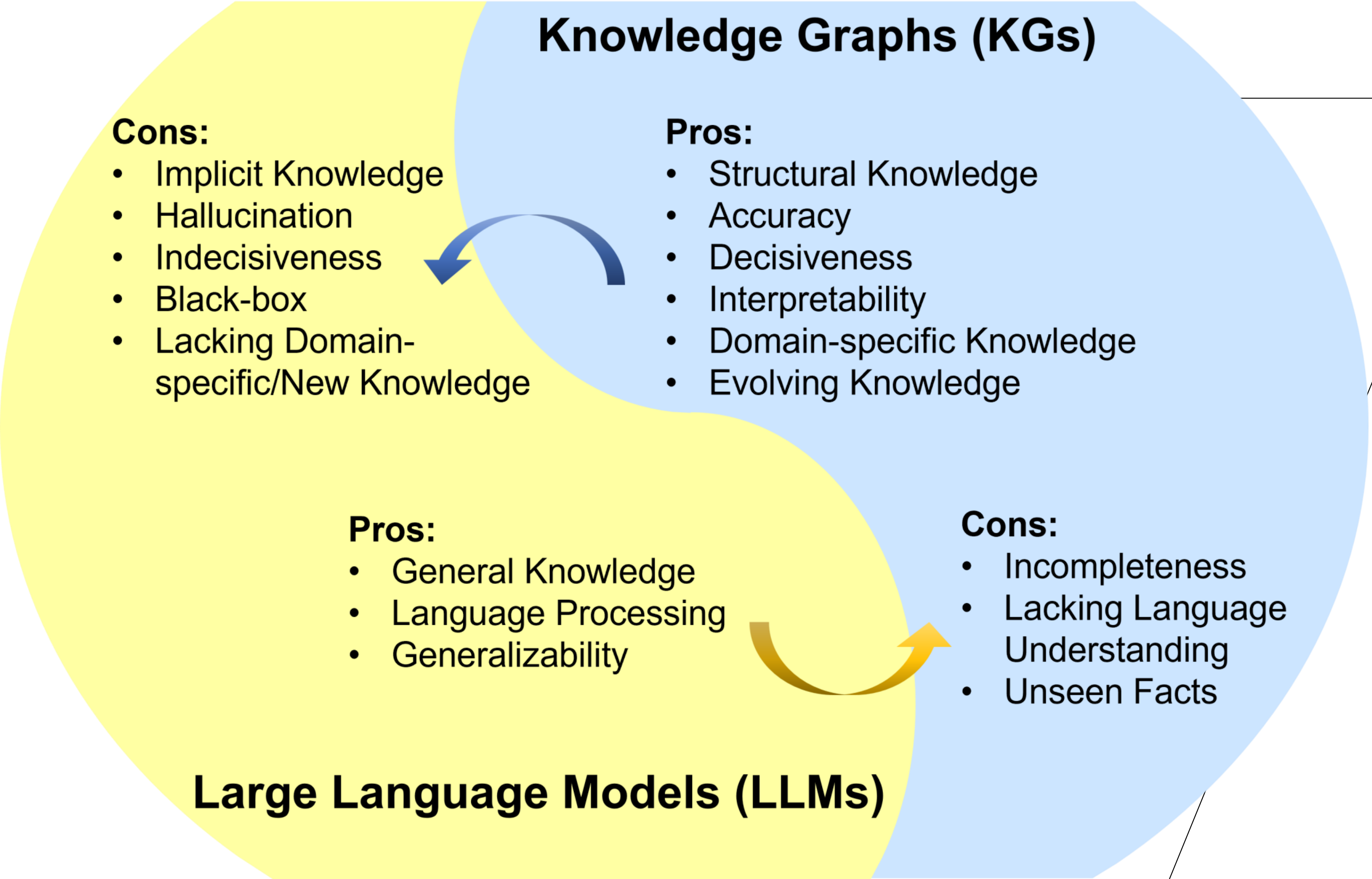
IYP: EXPLORATORY SEARCH



13 different datasets!

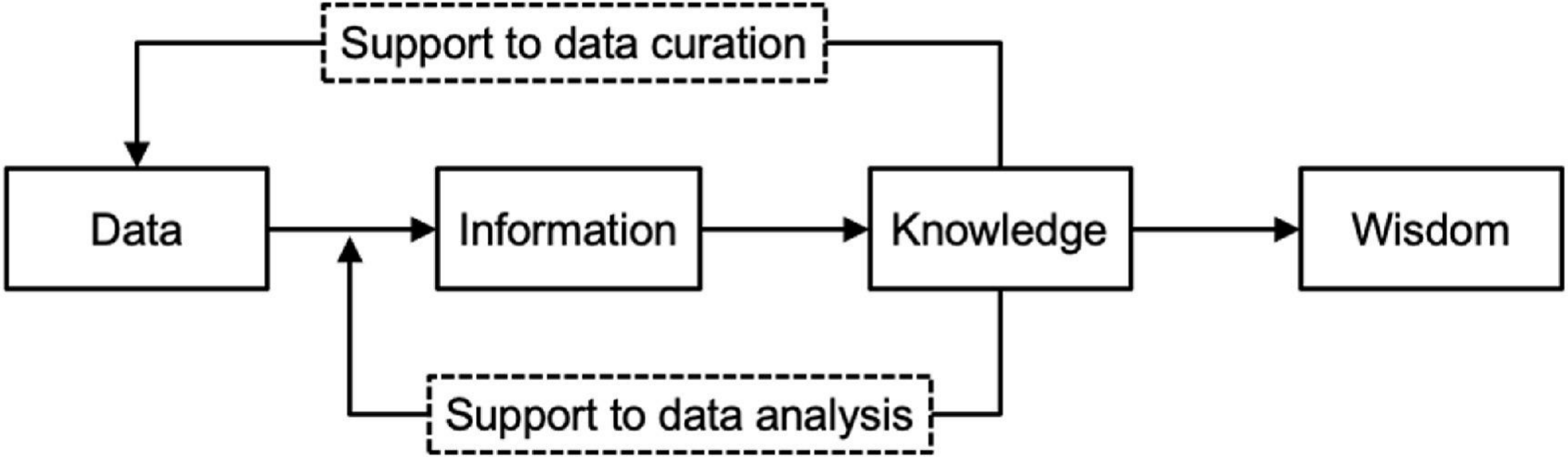
LLM VS. KG

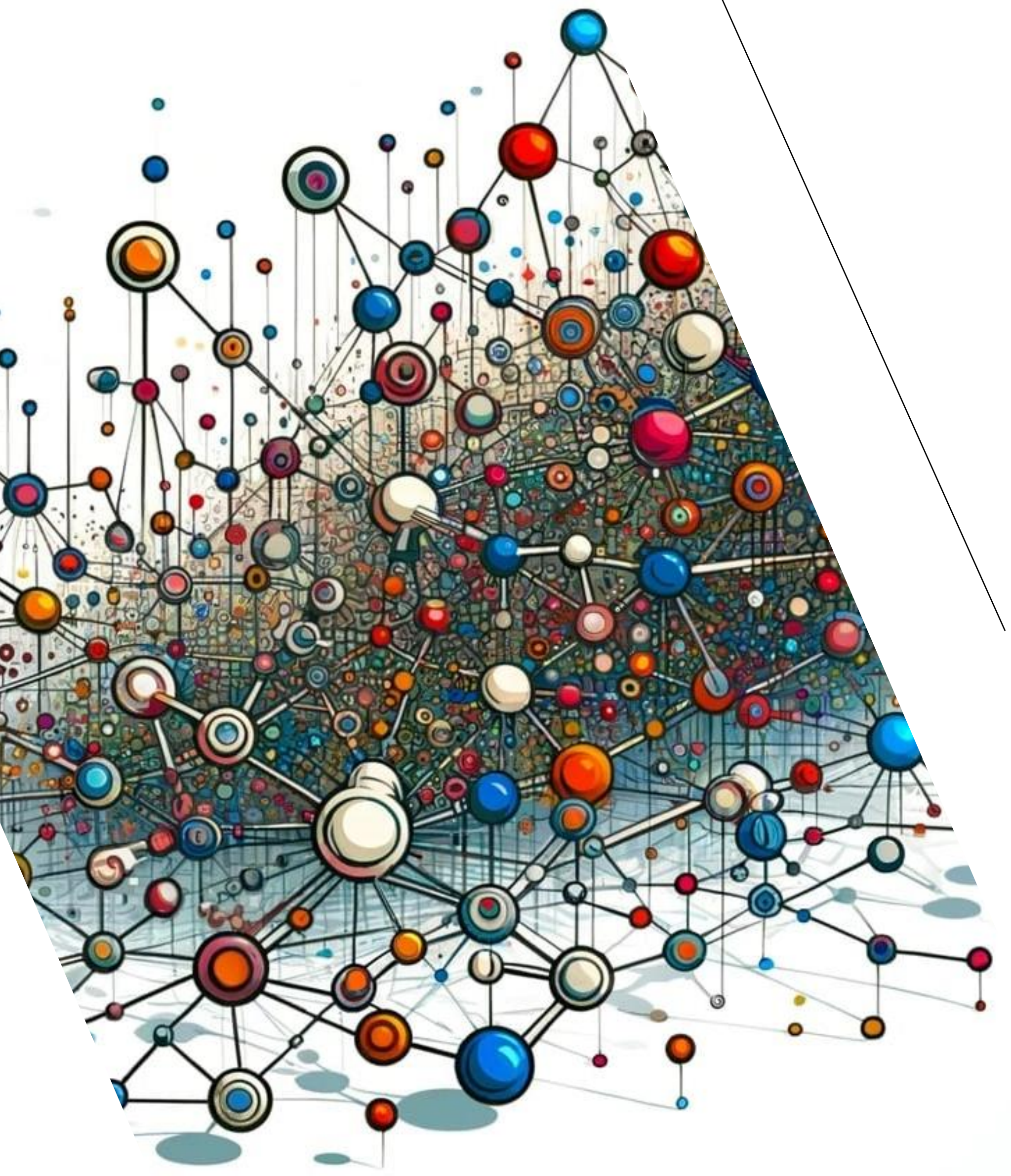
https://arxiv.org/html/2306.08302v3



THE ROLE OF MACHINE-READABLE KNOWLEDGE GRAPHS

https://www.sciencedirect.com/science/article/pii/S0098300422000450

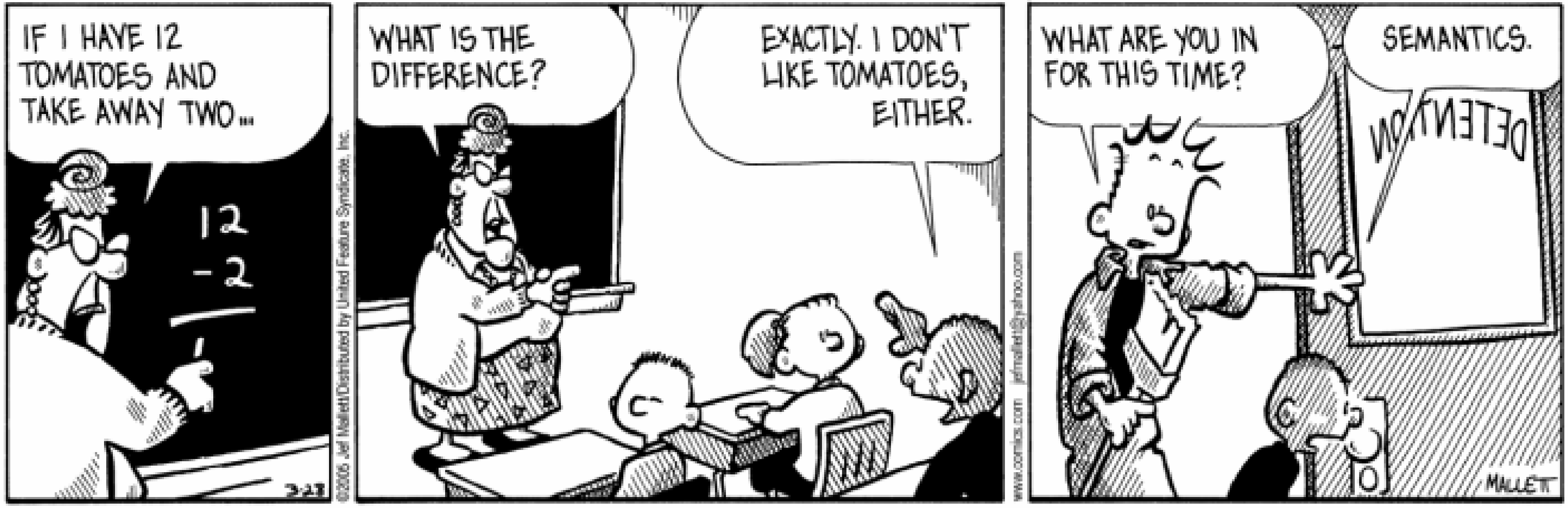




COMMON CHARACTERISTICS

- Large: Millions or billions of nodes & edges
- Coverage: Usually incomplete
- Correctness: how to resolve disagreeing datasets?
- Freshness: Depend on the kind of information

SEMANTICS


IF I HAVE 12 TOMATOES AND TAKE AWAY TWO...
12
-2
WHAT IS THE DIFFERENCE?
EXACTLY. I DON'T LIKE TOMATOES, EITHER.
WHAT ARE YOU IN FOR THIS TIME?
SEMANTICS.
MALLETT