UW xDD and COSMOS

Key updates

- PubMed abstract corpus now side-by-side with 13.4M full texts; API accessible, EMMAA/drugs pre-labeled
- Updated xdd-covid-19 full-content set
- Updated COSMOS pipeline deployed over sets, improvements to API (doc-level summary, scaled images)
- ASKE-ID: internal metadata tracking system for data and documents lacking DOIs
- Refined xDD “set” infrastructure

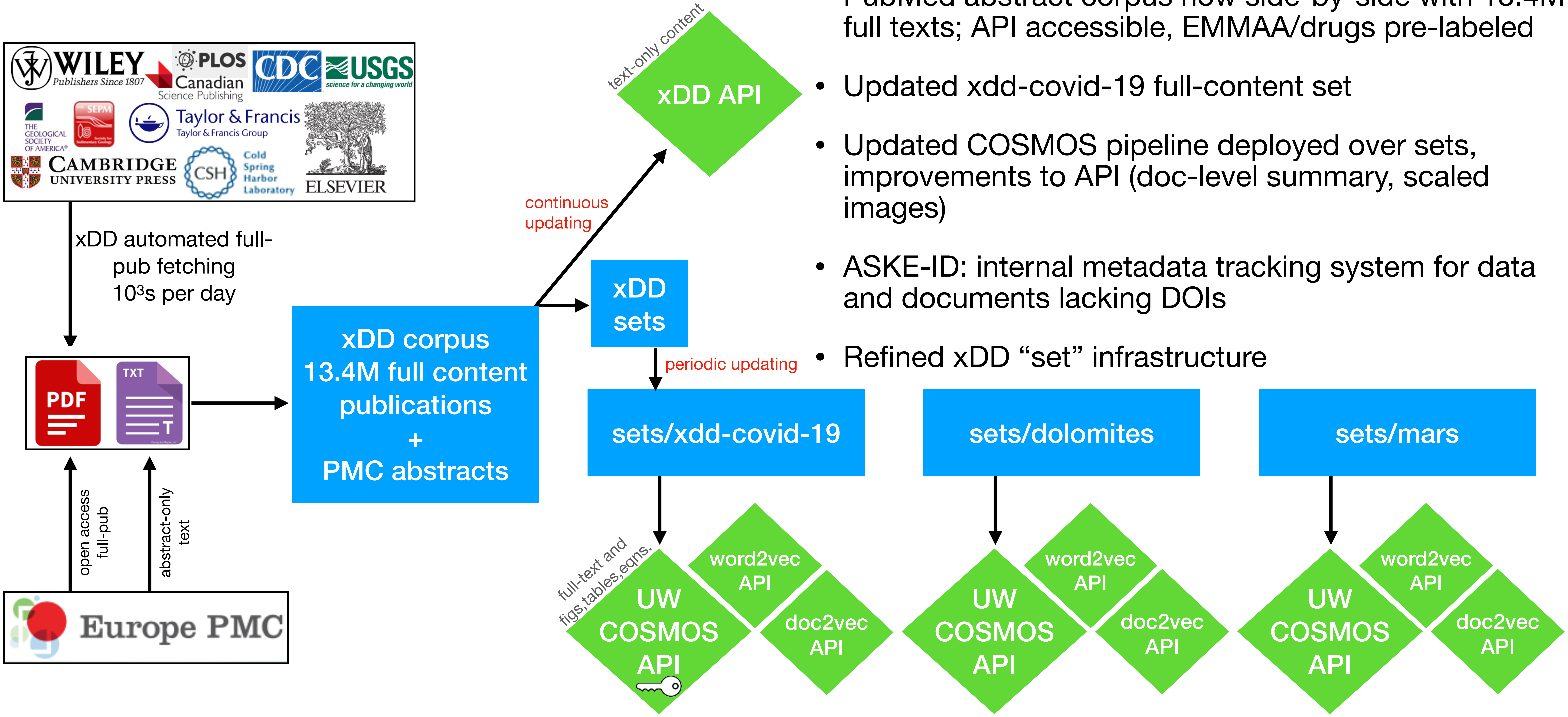


Table Co-References

Incorporate Body-Text Content into Table and Figure Retrieval

COSMOS | incidence rate | Table

Matthias Maneck, Christian Guenster, Hans-Joachim Meyer et al., *Influence of COVID-19 confinement measures on appendectomies in Germany - administrative claims data analysis of 9,797 patients*, **Cold Spring Harbor Laboratory Press, 2020**, DOI: 10.1101/2020.09.25.20198986

Table 1: Patient Demographics of hospital admissions with appendectomy in 2020 (ALL: all appendicitis stages; CAA: complex acute appendicitis; SAA: simple acute appendicitis; NAA: non acute appendicitis).

Week number	**Before 6-11**	**Lockdown 12-17**	**Easing 18-23**	**P**
ALL Cases	3,591	2,914	3,292	
Incidence rate per day	85.5	69.4	78.4	**<0.001**
Age in years				
Mean (sd)	32.7 (19.0)	34.5 (19.8)	34.9 (19.9)	**<0.001**
1-18 (%)	26.6	24.0	23.1	**0.002**
19-64 (%)	66.0	67.1	67.2	0.527
≥ 65 (%)	7.4	9.0	9.8	**0.002**
Female sex (%)	49.1	46.1	48.5	**0.042**
Appendicitis (%)				
Complex acute	21.2	27.0	23.6	**<0.001**
Simple acute	70.1	69.2	70.8	0.390
Non acute	8.6	3.8	5.7	**<0.001**
LOS, d; mean (sd)	4.3 (3.6)	4.3 (3.6)	4.3 (3.2)	0.567
CAA Cases	763	786	776	
Incidence rate per day	18.2	18.7	18.5	0.8373
Age in years				
Mean (sd)	42.4 (23.1)	41.9 (23.2)	44.2 (23.5)	0.118
1-18 (%)	21.9	21.4	19.1	0.348
19-64 (%)	59.4	59.9	59.0	0.935
≥ 65 (%)	18.7	18.7	21.9	0.191
Female sex (%)	41.0	41.2	43.3	0.603
LOS, d; mean (sd)	6.9 (5.2)	6.8 (5.0)	6.9 (4.7)	0.844
SAA Cases	2,518	2,016	2,330	
Incidence rate per day	60.0	48.0	55.5	**<0.001**
Age in years				
Mean (sd)	30.1 (17.0)	31.8 (17.7)	32.2 (17.8)	**<0.001**
1-18 (%)	27.7	25.2	24.4	**0.026**
19-64 (%)	67.6	69.1	69.4	0.371
≥ 65 (%)	4.7	5.6	6.2	0.061
Female sex (%)	49.0	46.9	49.0	0.273
LOS, d; mean (sd)	3.6 (2.5)	3.4 (2.1)	3.5 (1.8)	**<0.001**
NAA Cases	310	112	186	

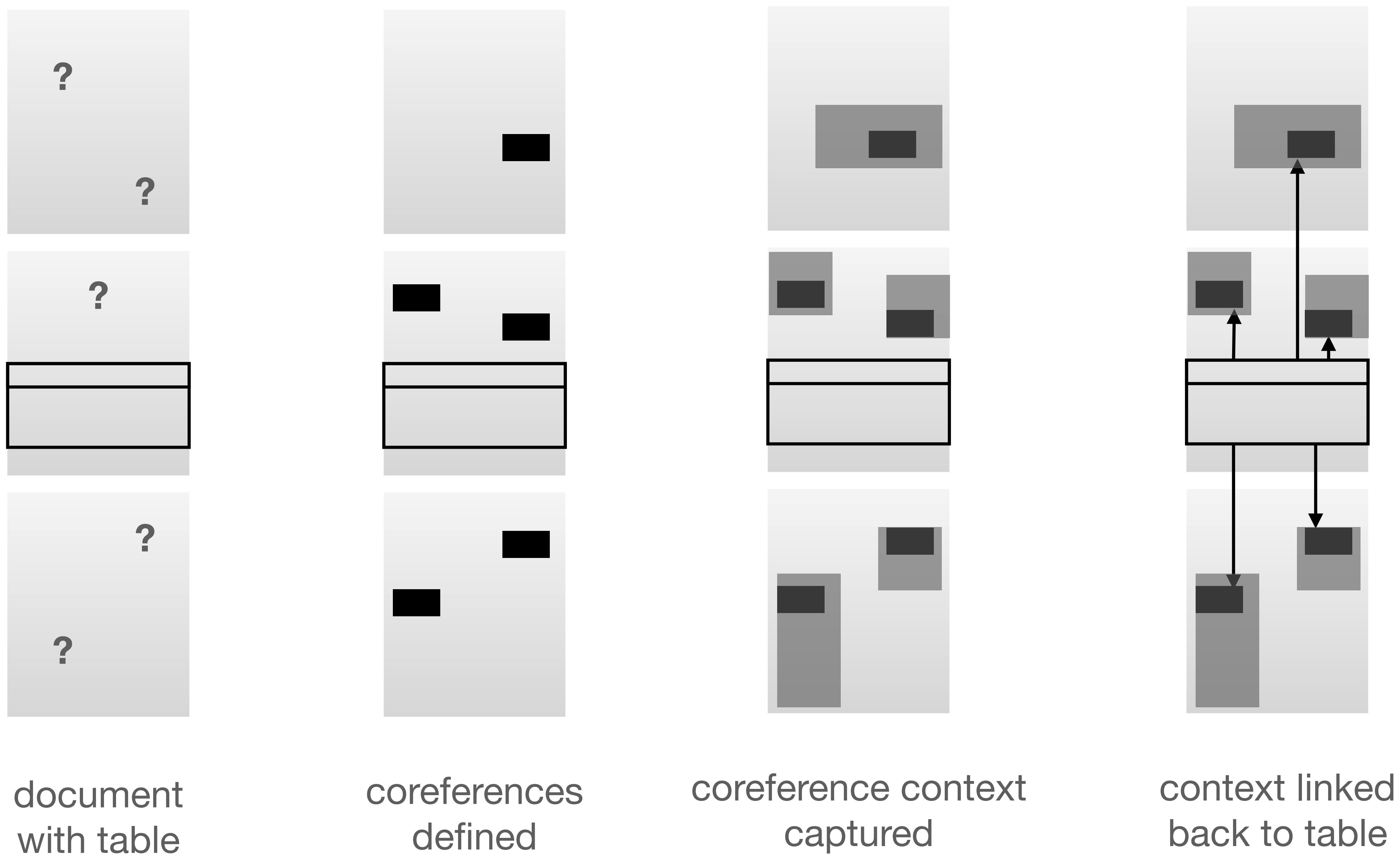


Table Co-References

Incorporate Body-Text Content into Table and Figure Retrieval

Table Reference

Table Caption

TABLE. Hospitalization, intensive care unit (ICU) admission, and case–fatality percentages for reported COVID–19 cases, by age group — United States, February 12–March 16, 2020

Table

Age group (yrs) (no. of cases)	%*		
	Hospitalization	ICU admission	Case-fatality
0–19 (123)	1.6–2.5	0	0
20–44 (705)	14.3–20.8	2.0–4.2	0.1–0.2
45–54 (429)	21.2–28.3	5.4–10.4	0.5–0.8
55–64 (429)	20.5–30.1	4.7–11.2	1.4–2.6
65–74 (409)	28.6–43.5	8.1–18.8	2.7–4.9
75–84 (210)	30.5–58.7	10.5–31.0	4.3–10.5
≥85 (144)	31.3–70.3	6.3–29.0	10.4–27.3
Total (2,449)	20.7–31.4	4.9–11.5	1.8–3.4

Table Note

* Lower bound of range = number of persons hospitalized, admitted to ICU, or who died among total in age group; upper bound of range = number of persons hospitalized, admitted to ICU, or who died among total in age group with known hospitalization status, ICU admission status, or death.

Ref: CDC COVID-19 Response Team. Severe outcomes among patients with coronavirus disease 2019 (COVID-19)—United States, February 12–March 16, 2020. MMWR Morb Mortal Wkly Rep 2020;69:343–6. 10.15585/mmwr.mm6912e2

Table Reference Context

aged ≤19 years. Percentages of ICU admissions were lowest among adults aged 20–44 years (2%–4%) and highest among adults aged 75–84 years (11%–31%) (Table).

Table Reference

Among 44 cases with known outcome, 15 (34%) deaths were reported among adults aged ≥85 years, 20 (46%) among adults aged 65–84 years, and nine (20%) among adults aged 20–64 years. Case-fatality percentages increased with increasing age, from no deaths reported among persons aged ≤19 years to highest percentages (10%–27%) among adults aged ≥85 years (Table) (Figure 2).

Table Reference Context

Table Reference

Among 508 (12%) patients known to have been hospitalized, 9% were aged ≥85 years, 36% were aged 65–84 years, 17% were aged 55–64 years, 18% were 45–54 years, and 20% were aged 20–44 years. Less than 1% of hospitalizations were among persons aged ≤19 years (Figure 2). The percentage of persons hospitalized increased with age, from 2%–3% among persons aged ≤9 years, to ≥31% among adults aged ≥85 years. (Table).

Table Reference Context

Table Reference

Adults aged 20-44

Increasing age

Persons hospitalized

TABLE. Hospitalization, intensive care unit (ICU) admission, and case–fatality percentages for reported COVID–19 cases, by age group — United States, February 12–March 16, 2020

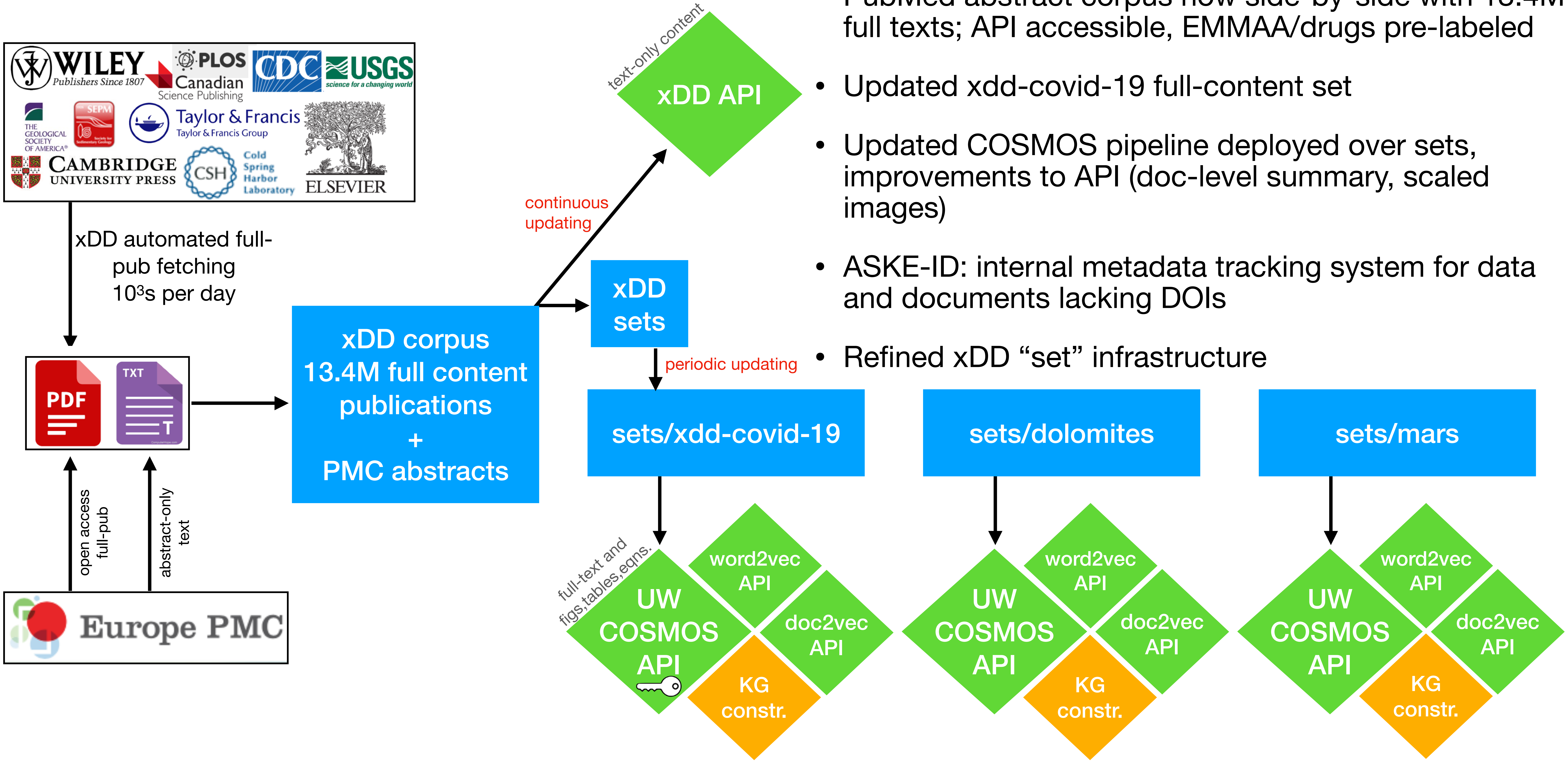
Age group (yrs) (no. of cases)	%*		
	Hospitalization	ICU admission	Case-fatality
0–19 (123)	1.6–2.5	0	0
20–44 (705)	14.3–20.8	2.0–4.2	0.1–0.2
45–54 (429)	21.2–28.3	5.4–10.4	0.5–0.8
55–64 (429)	20.5–30.1	4.7–11.2	1.4–2.6
65–74 (409)	28.6–43.5	8.1–18.8	2.7–4.9
75–84 (210)	30.5–58.7	10.5–31.0	4.3–10.5
≥85 (144)	31.3–70.3	6.3–29.0	10.4–27.3
Total (2,449)	20.7–31.4	4.9–11.5	1.8–3.4

* Lower bound of range = number of persons hospitalized, admitted to ICU, or who died among total in age group; upper bound of range = number of persons hospitalized, admitted to ICU, or who died among total in age group with known hospitalization status, ICU admission status, or death.

- Table/Figure retrieval via API will be (optionally) enhanced by including additional body text context
- Embedding models trained over table/figure text will include enhanced content

UW xDD and COSMOS

Key updates



- PubMed abstract corpus now side-by-side with 13.4M full texts; API accessible, EMMAA/drugs pre-labeled
- Updated xdd-covid-19 full-content set
- Updated COSMOS pipeline deployed over sets, improvements to API (doc-level summary, scaled images)
- ASKE-ID: internal metadata tracking system for data and documents lacking DOIs
- Refined xDD “set” infrastructure

marius

A no-code system for learning massive graph embeddings on a single machine

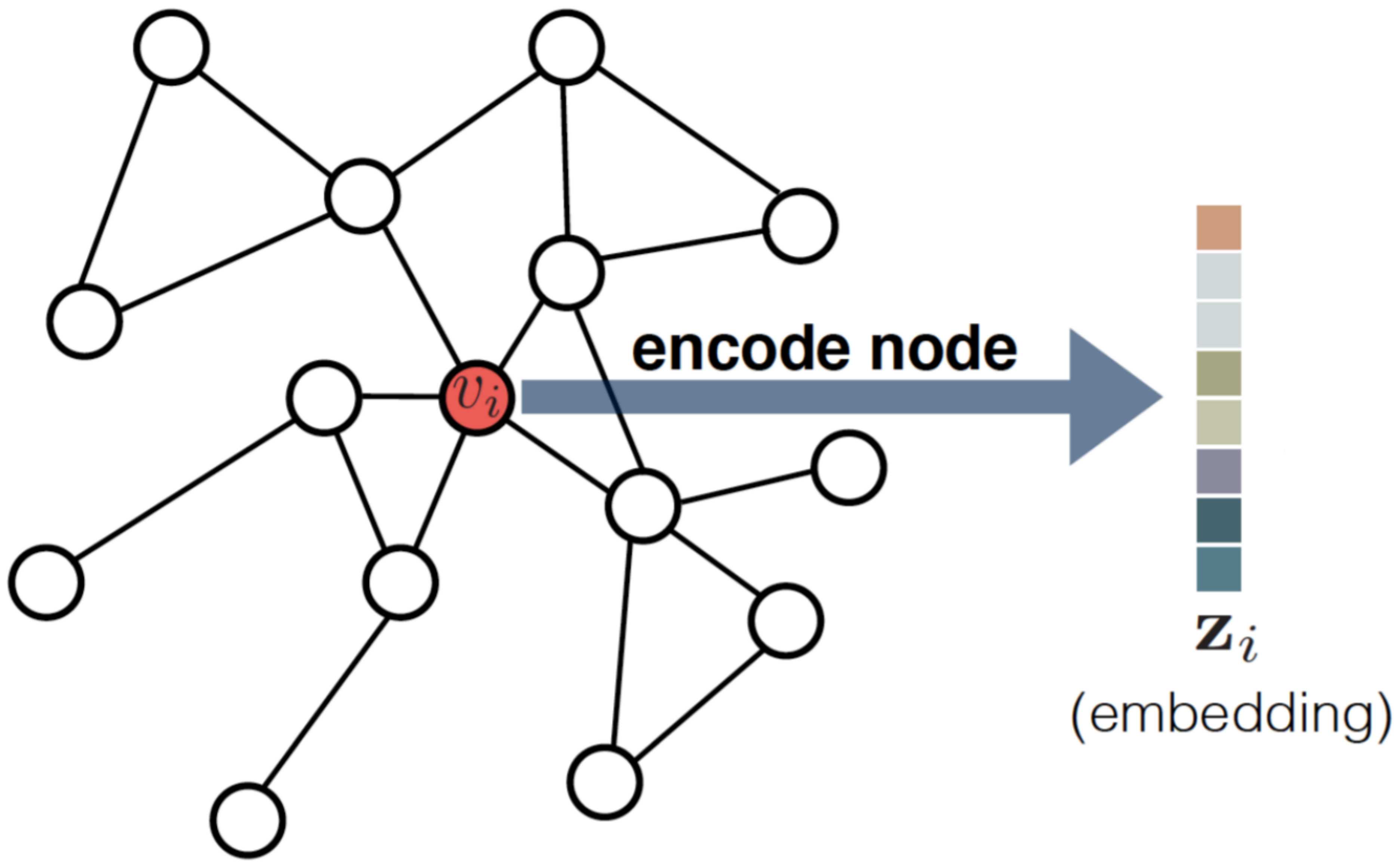
www.marius-project.org

Project leads: Theo Rekatsinas, Shivaram Venkataraman

Refs:

- **Marius: Learning Massive Graph Embeddings on a Single Machine**, Jason Mohoney, Roger Waleffe, Shanan Peters, Theodoros Rekatsinas, Shivaram Venkataraman, To Appear in OSDI 2021
- **Demo of Marius: Graph Embeddings with a Single Machine**, Anders Carlsson, Anze Xie, Jason Mohoney, Roger Waleffe, Shanan Peters, Theodoros Rekatsinas, Shivaram Venkataraman, Under Submission VLDB 2021

Graph Embeddings

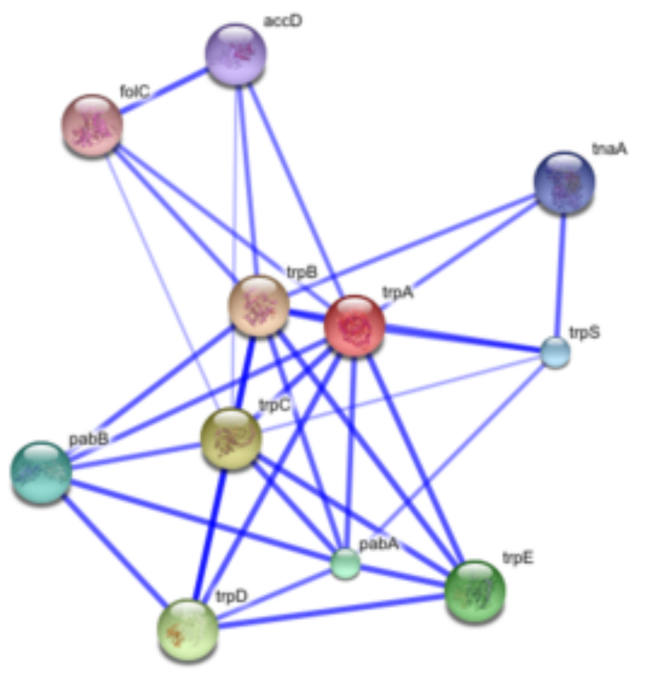


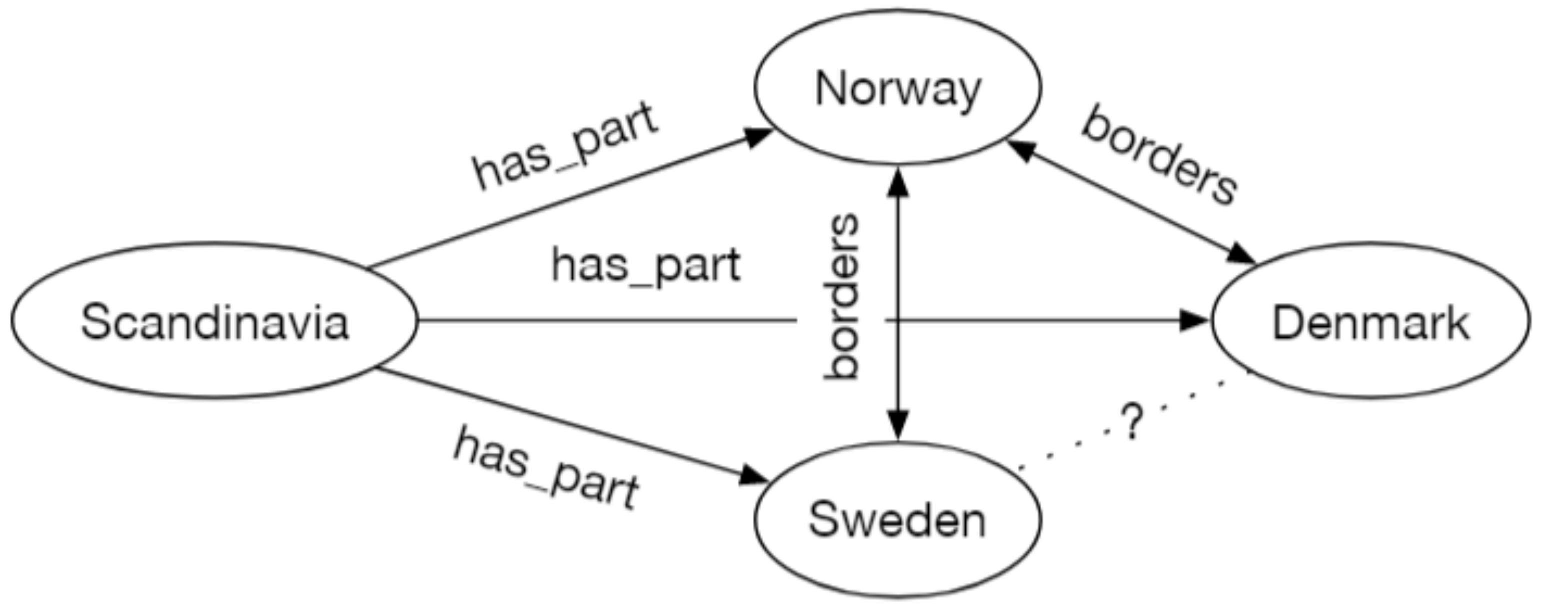
Graph embedding is the process of converting discrete graph representations into continuous vector representations which preserve the property of the original graph structure

Graph Embedding Examples

Link prediction: Predict potential connections in a social network

Node classification: Predict protein functions based on related proteins





Relation inference: Predict missing edges in a knowledge graph marius

Challenge: Massive Data Volume

- Large-scale graph embedding is memory intensive

  E.g. Wikidata dataset: 92.9 million nodes, embedding dimension 400

  Requires 148 GB to store embeddings of all nodes

- State-of-the-art methods distribute training across multiple machines and require large amount of computing resources

PyTorch-BigGraph from Facebook

DGL-KE from Amazon

Our solution: Marius

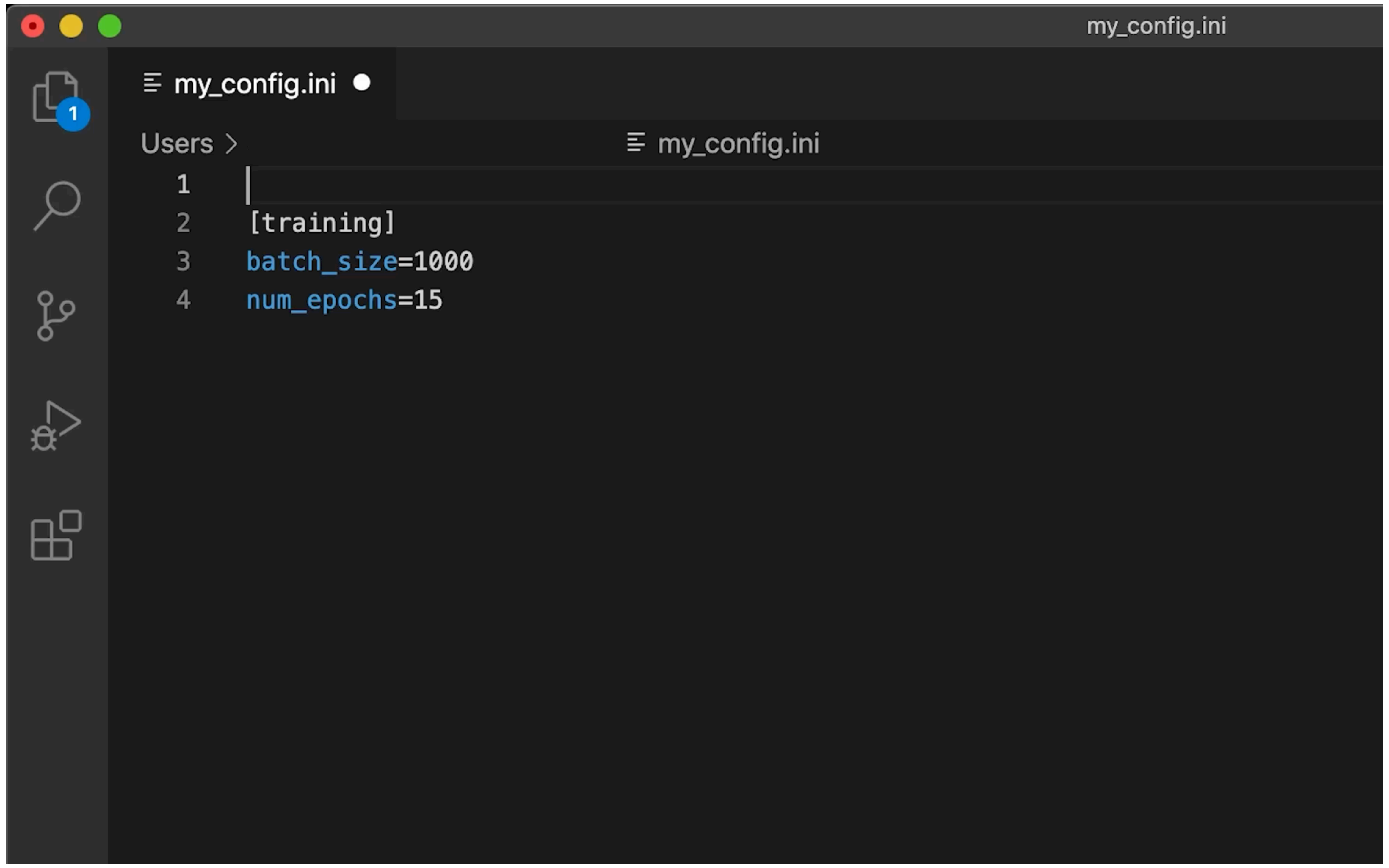
- Allows large-scale graph embedding on a single machine
- Easy-to-use config-based development framework
- An open-source system introduced in OSDI 2021

Config Based

- No-code paradigm: running Marius only requires a simple configuration file

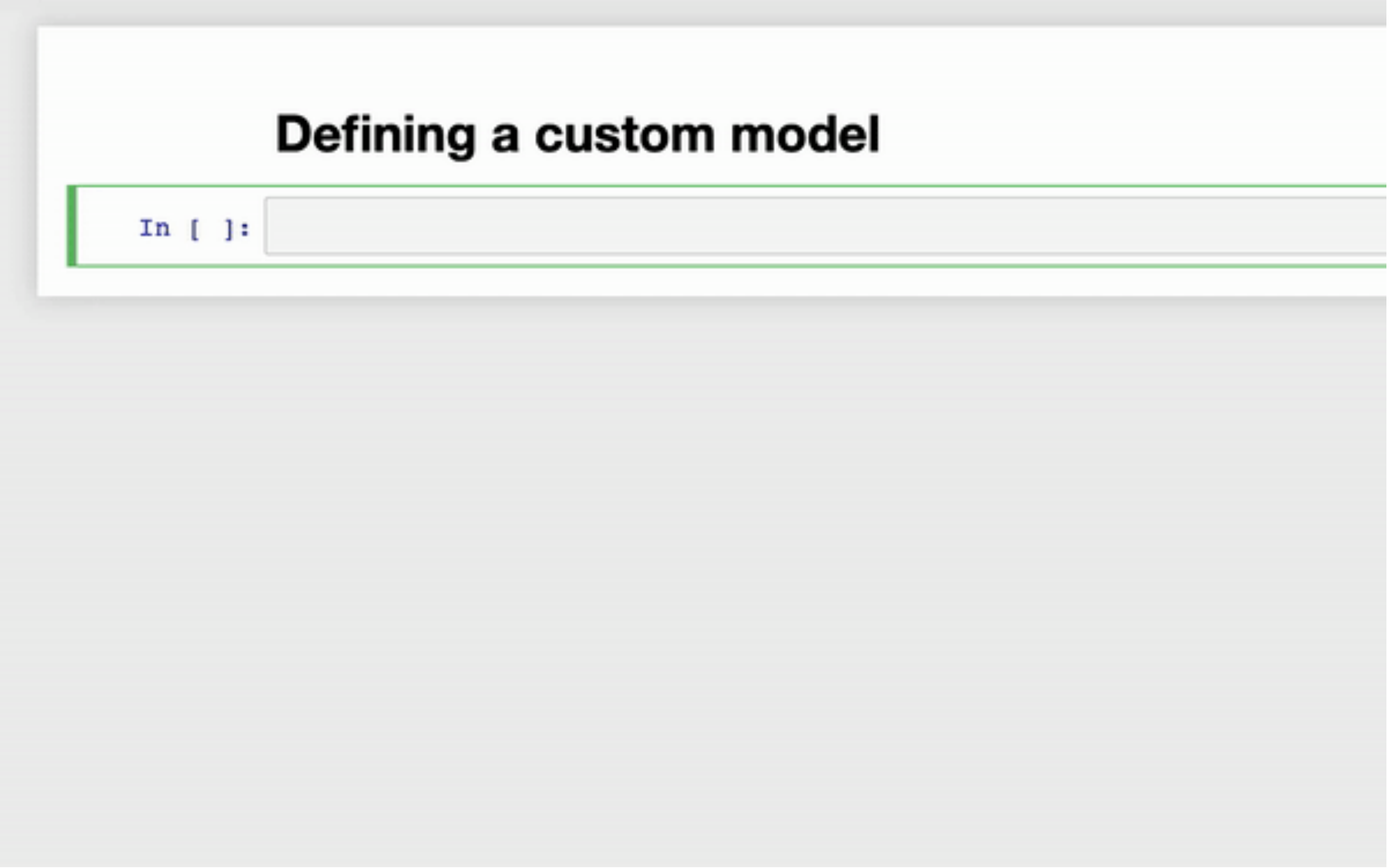
- Customize parameters, defaults provided if not specified

- Easily run from command line



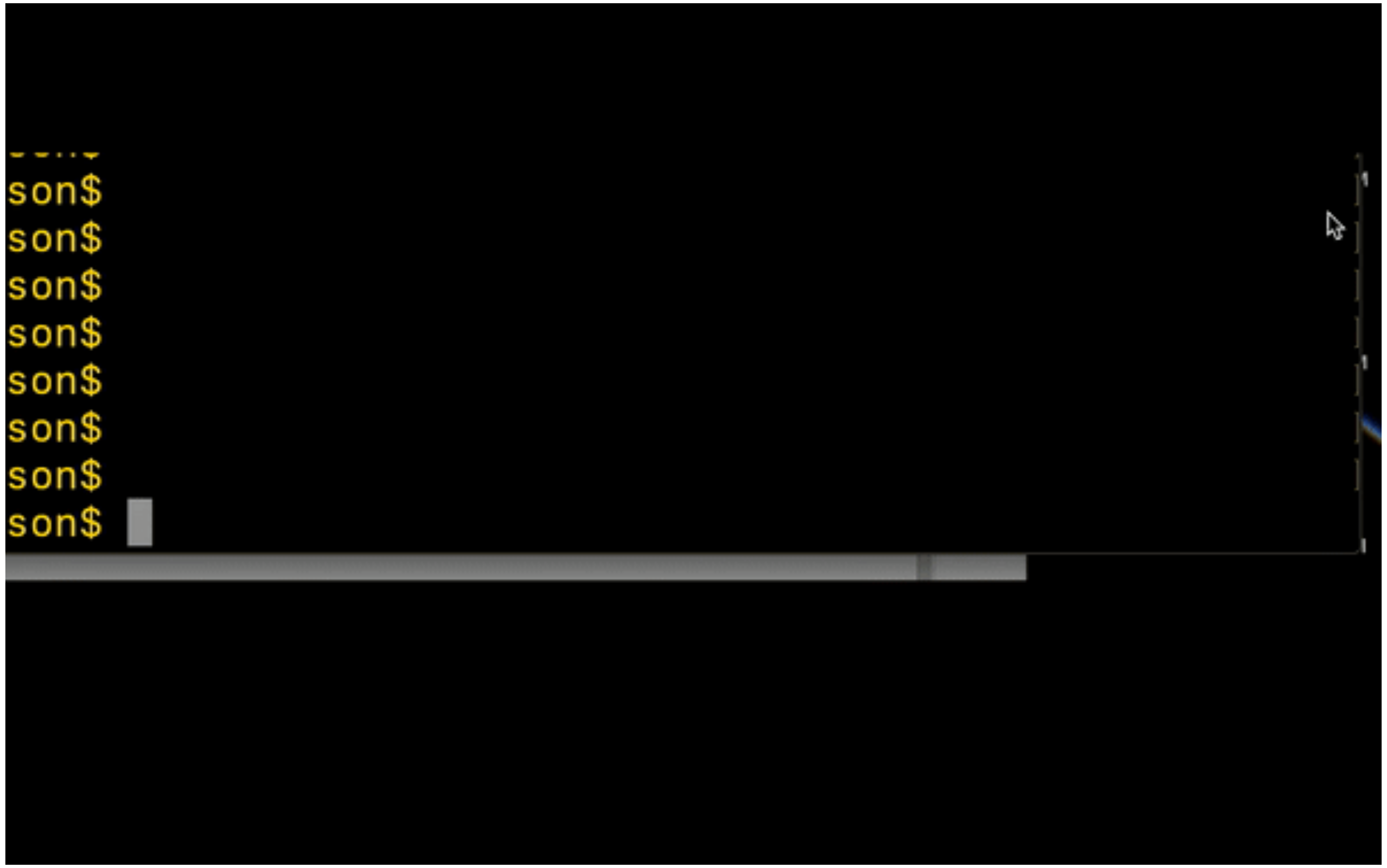
Extend Marius with Python

- Features Python API
- Write custom models
- Allows a high degree of control and customization



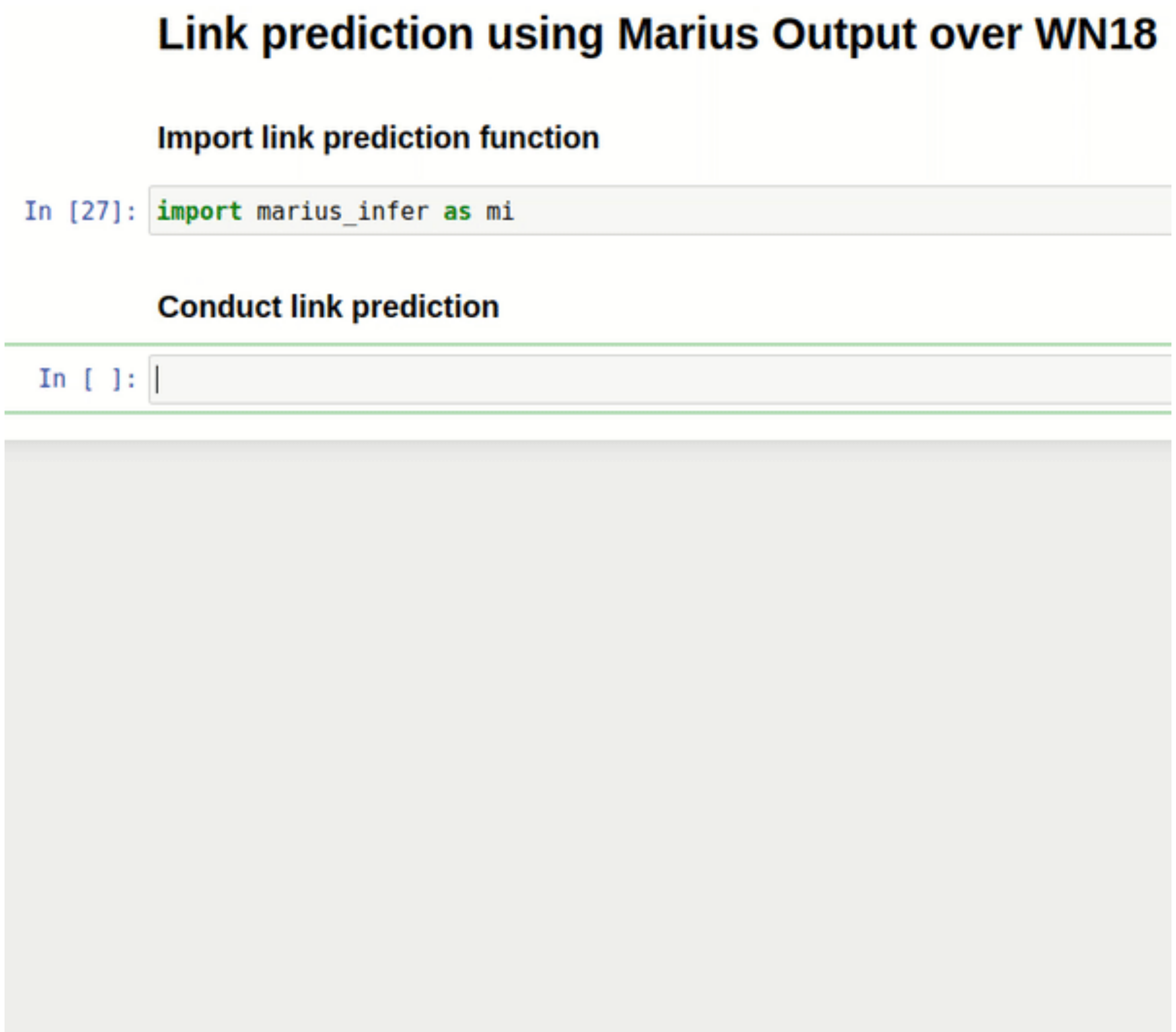
Data Converters

- Transforms raw data into the Marius input format
- Support for conversion of TSV, CSV, Parquet file formats
- Can be invoked by a single command or from Python
- 31 datasets supported out of the box
- Output embeddings can be converted to commonly used types such as PyTorch tensors

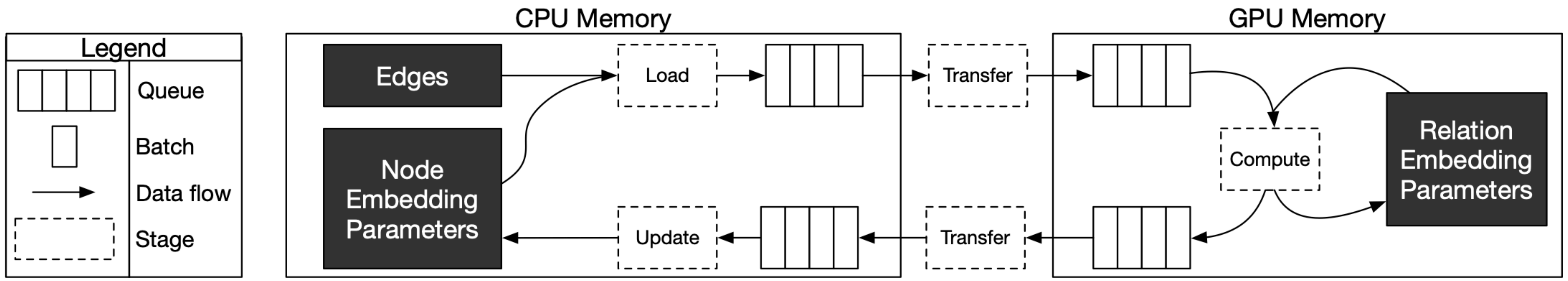


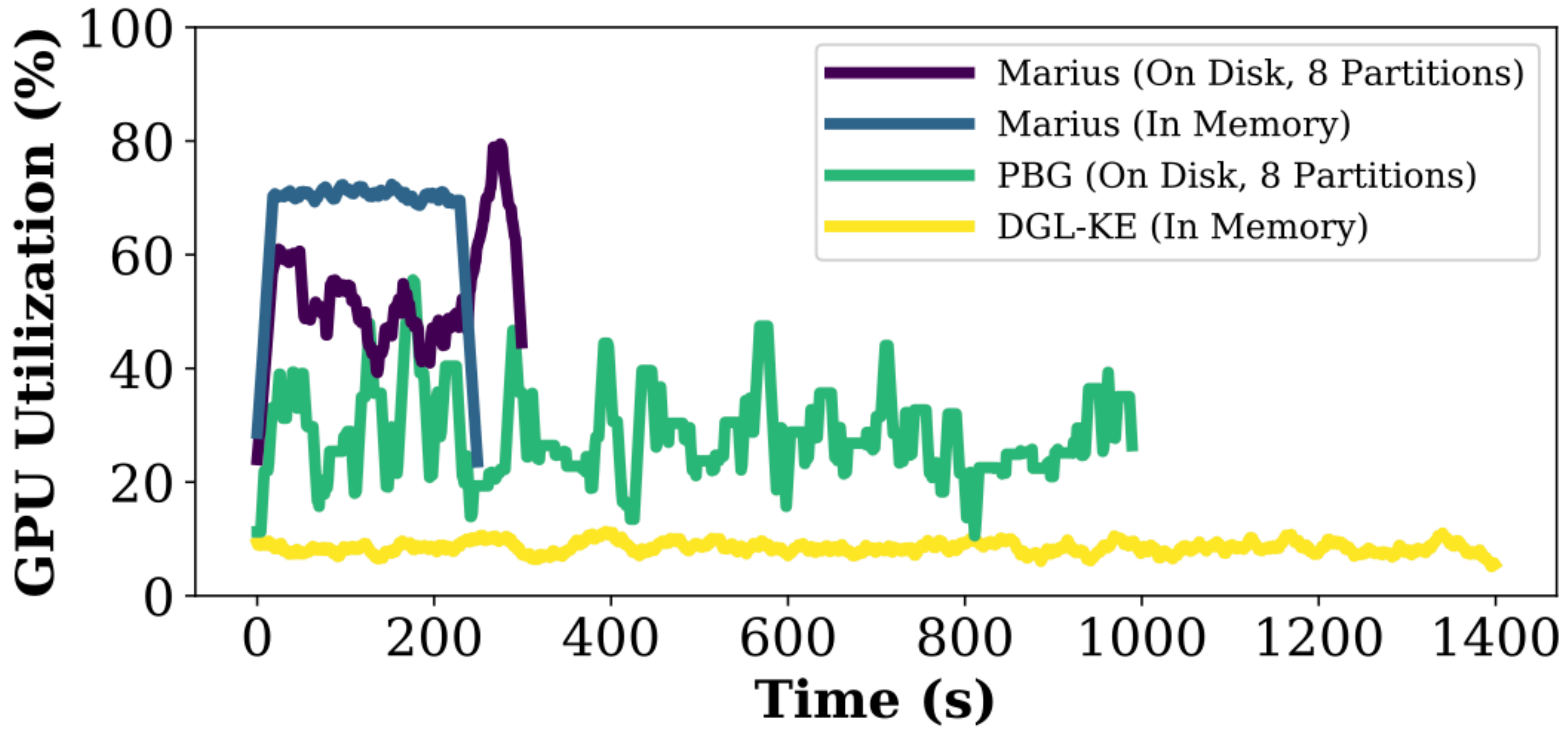
Inference with Marius Output for WN18

- Postprocessor methods are provided to transform embeddings to usable formats such as PyTorch tensors
- Downstream inference tasks such as link prediction can be conducted easily
- Query results are accurate and meaningful



Scalable graph learning



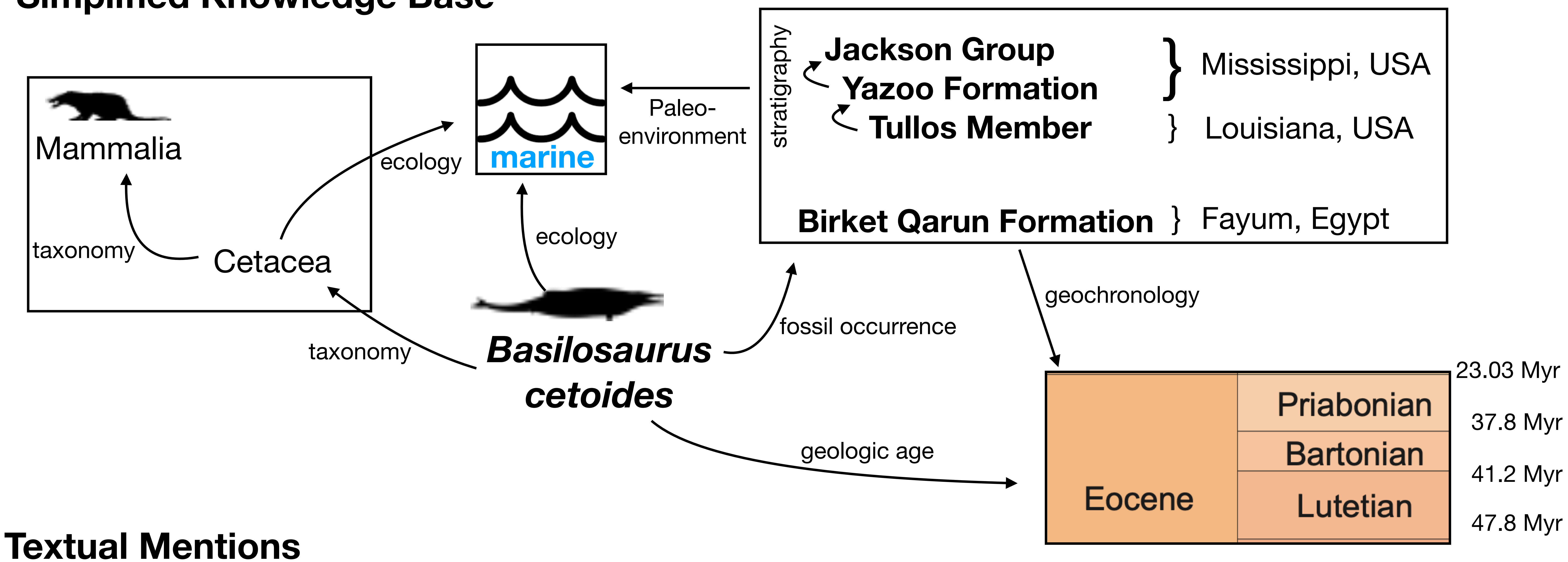


Pipelining and a novel data replacement policy allow Marius to maximize resource utilization of the entire memory hierarchy (including disk, CPU, and GPU memory)

Achieves graph learning over billion edge graphs **in a single machine**

Integrating Marius with COSMOS

Simplified Knowledge Base



Analogical Reasoning Example

Basilosaurus —geochronology→ **Eocene**

Pakicetus —geochronology→ **?**

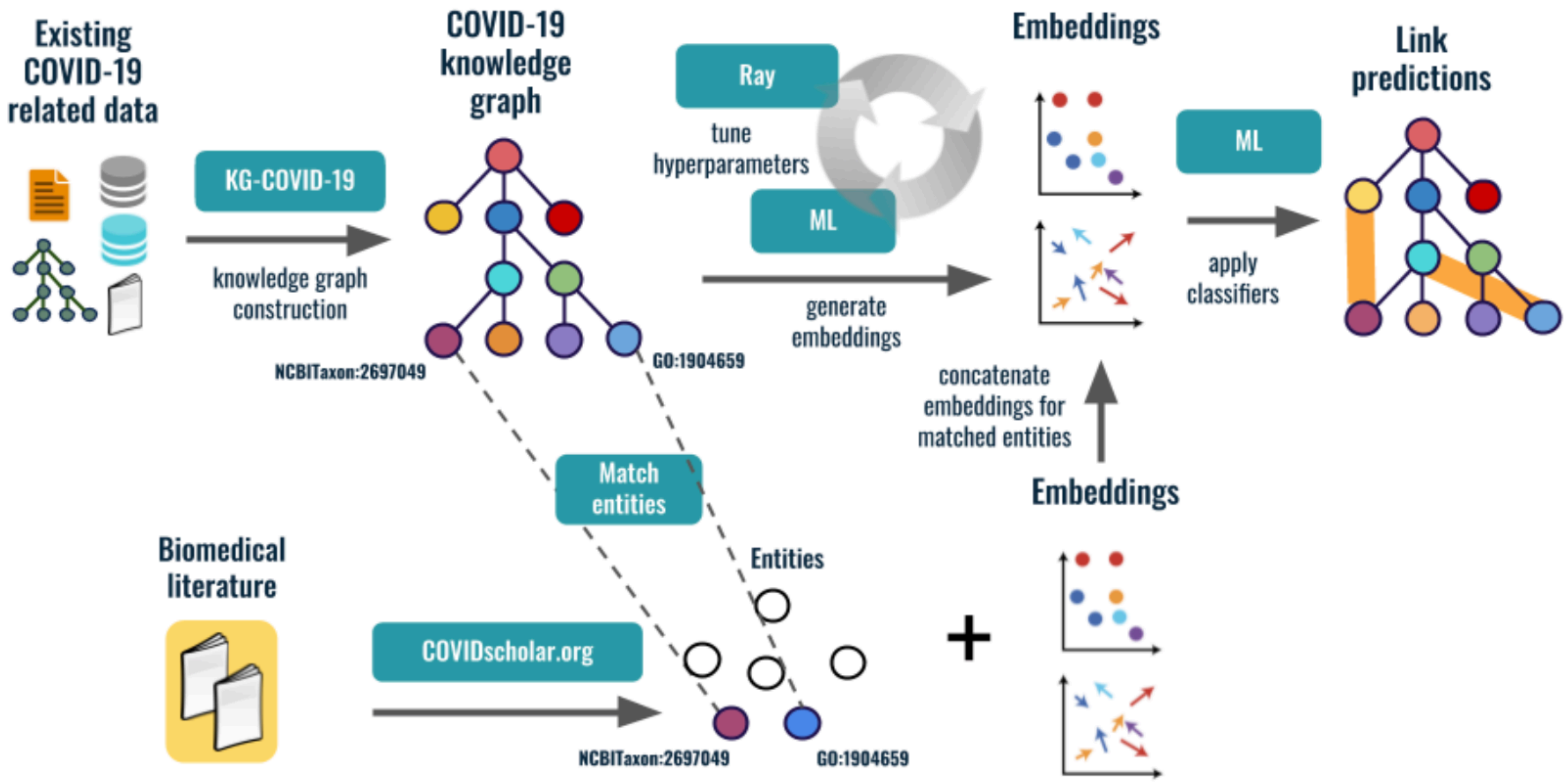
Textual Mentions

Fossils from an extinct toothed (**Archaeocete**) **whale**, ***Basilosaurus cetoides***, were found in a surface exposure of the **Pachuta Marl Member** of the **late Eocene Yazoo Clay** near the Matherville community in **Wayne County, Mississippi**.

The **Yazoo Clay Formation** makes up the upper half of the **Jackson Group** in the central **Gulf Coastal Plain**, representing deposition during the TAGC4.3 **marine** transgression.

Joint embeddings of text and existing knowledge graphs to enable analogical reasoning and knowledge completion in any domain

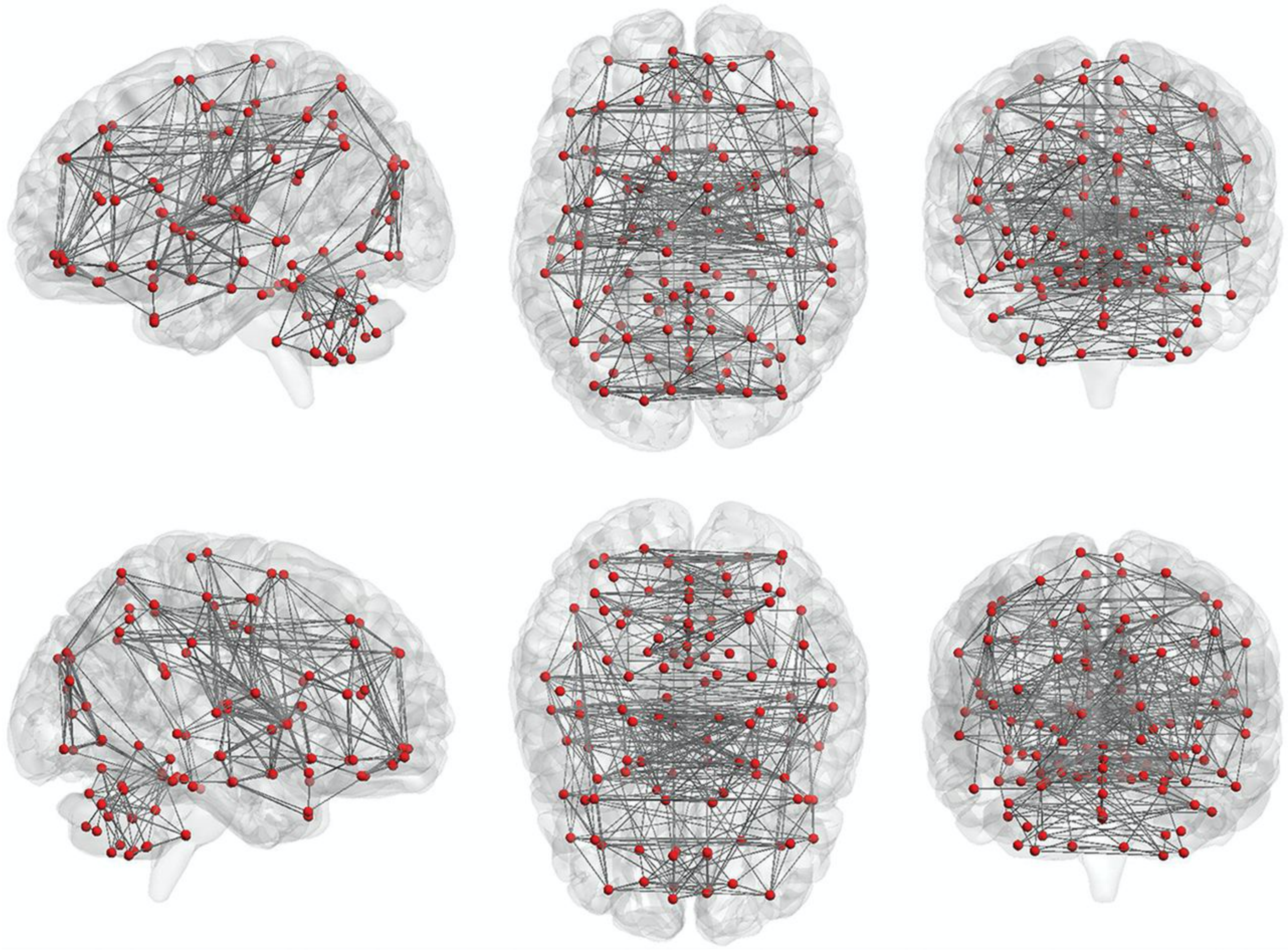
Integrating Marius with COSMOS



Instead of learning individual embeddings and then concatenate; use Marius to learn a joint geometric representation

Source: http://berkeleybop.org/project/kg-covid-19/

Marius beyond COSMOS



Model dependencies and interactions between signal measurements

Source: http://www.ajnr.org/content/early/2018/01/18/ajnr.A5527