

What is the role of data management in the age of AI?

Chris Bizon
Renaissance Computing Institute
University of North Carolina, Chapel Hill

May 4, 2025 - me to my family

Bulls game is on TV and they assigned a ball to the pitcher because he didnt make eye contact with the batter. Was confused about the rule so I searched it on google. The google AI starts strong but the answer took a somewhat unusual turn!

2:16 PM · MMS

May 4, 2025 - AI to me



Search Labs | AI Overview

In Major League Baseball, the new pitch clock rules emphasize the importance of eye contact between the pitcher and batter. The rule is **designed to reduce game time and prevent stalling**. Making eye contact signals that the pitcher is ready to throw, and the batter is expected to be in the box and attentive. Refusing to make eye contact can result in a "D-sist," and after five such instances, the pitcher can challenge the hitter to a duel. [🔗](#)



New York Times, 2023



The New York Times

<https://www.nytimes.com> › athletic › 2023/06/14 › glan... ⋮

Doug Glanville: MLB's new rules have turned umpires into ...

Jul 26, 2023 — ... **pitcher** or hitter **refuses** to **make eye contact** and gets a warning for it. **After five** Disses, a **pitcher** may **challenge** the hitter to a **duel**.

Garbage In, Garbage Out ?

- Not really ... the original article just contains a joke
- Even if the joke (or the opinion) is bad, it's not technically garbage
- However google is pulling in this information is not able to understand its context and appropriate use
- Maybe an LLM should be able to tell the difference, maybe not, but there's also *no metadata* here to help.

What if you want your bot to respond with facts?

- Have an agent look at the question and find relevant facts
- Feed these facts to the LLM along with the user question to generate a response.
- You're using the LLM for what it's good at: text processing, and (hopefully) not what it's only so-so at: knowing things and reasoning
- (Caveat that might be changing over time)
- RAG: Retrieval Augmented Generation

It's up to the RAG-builder to populate this bot with the corpus high-quality documents that the bot will serve!

Put some Triples together, that's a Knowledge Graph

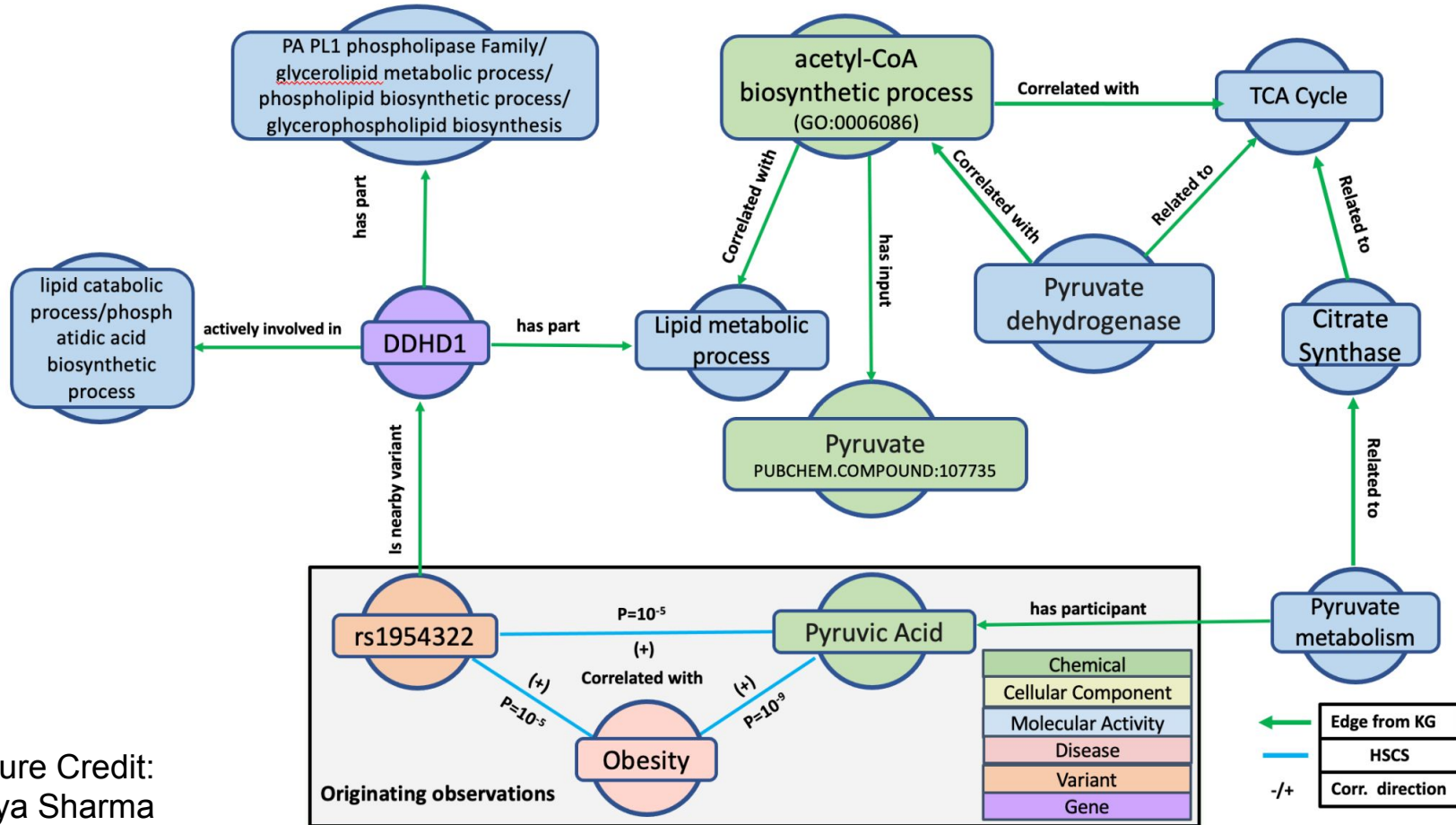
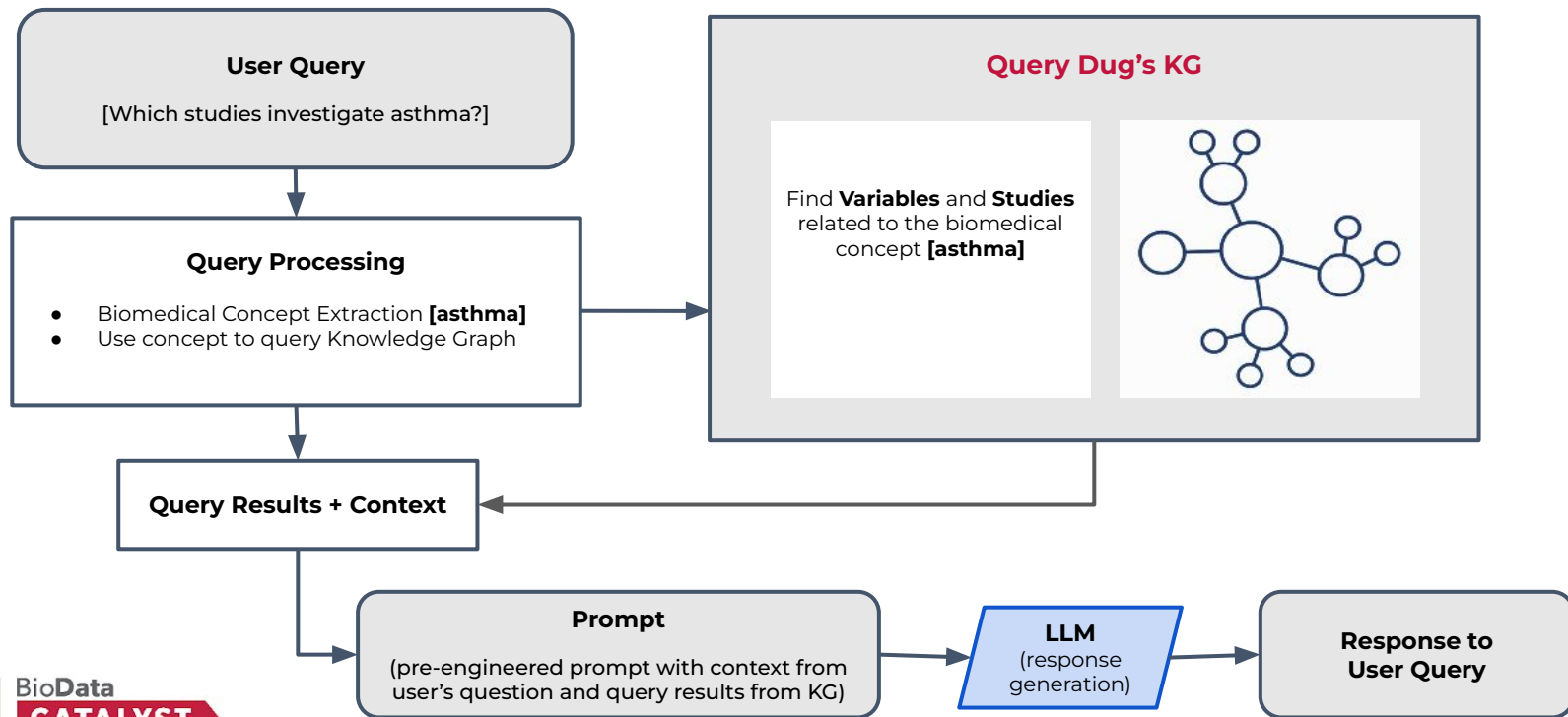


Figure Credit:
Priya Sharma

DugBot Knowledge Graph Agent

Courtesy Suparna Goswami (RENCI)

- Query Dug's knowledge Graph using biomedical concepts



There's more to AI than LLMs

ChatGPT o3-mini-high ▾

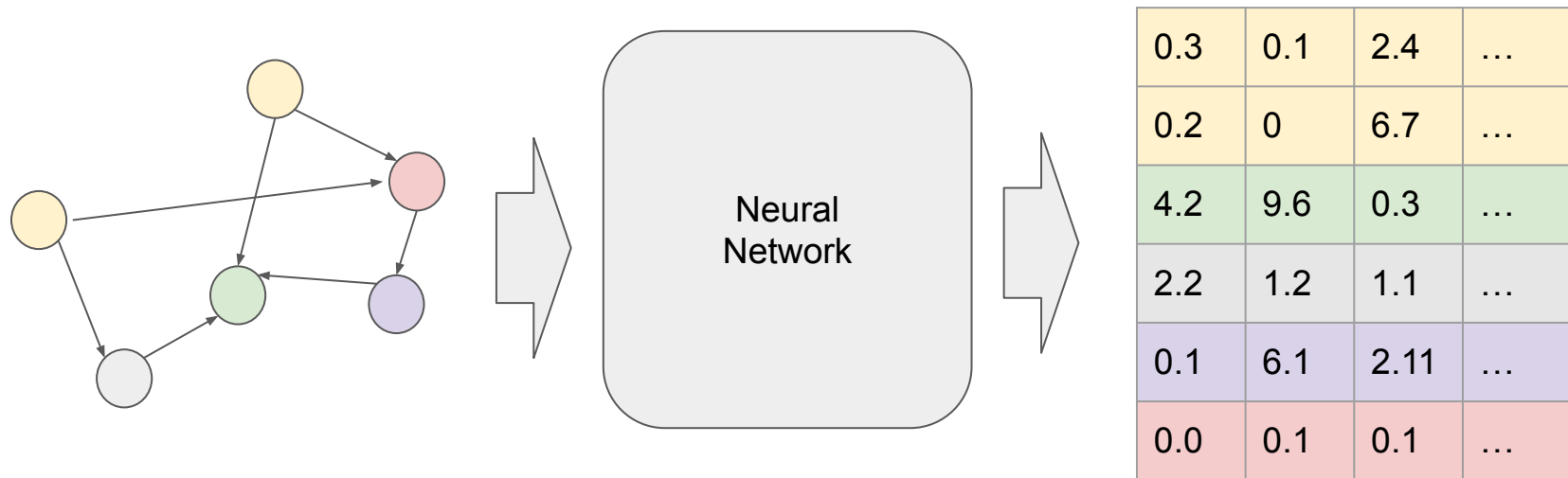
Analyze every possible combination of FDA approved drugs and known diseases to suggest the top 10 most likely drug repurposing opportunities

Reasoned for 44 seconds >

Below is a conceptual synthesis—drawing on systems pharmacology, transcriptomics, network analysis, and literature evidence—that outlines a set of ten promising repurposing hypotheses. In practice, “analyzing every possible combination” of FDA-approved drugs versus all known diseases is an enormous computational task. Instead, researchers typically integrate multi-omics data, drug–target networks, and epidemiological insights to highlight candidates with strong mechanistic rationale. The list below represents candidates that recurrently emerge from such approaches. (Keep in mind that these hypotheses require further experimental and clinical validation.)

Courtesy A. Tropsha
(UNC)

Embeddings turn topology into geometry (points in space)



Embeddings can be optimized in different ways, reflecting different assumptions about the embedding space.

This is analogous to the way that large language models embed words

We can derive answers, but where do they come from?

This system is for research purposes and is not meant to be used by clinical service providers in the course of treating patients.



What drugs may treat conditions related to
Immune-mediated Necrotizing Myopathy?

Share Result Set

Results

Showing 1-10 of 510 Results

1

2

3

4

...

51

>

Filters

NAME

EVIDENCE

PATHS ⓘ

SCORE ⓘ ↓

Mycophenolate Mofetil



Publications (92)
Clinical Trials (4)
Sources (7)

14 Paths

5.00



Immunosuppressive Agent (18)

Pro-agent (19)

Ec 1.* (oxidoreductase) Inhibitor (40)

Antimicrobial Agent (32)

The morpholinoethyl ester of mycophenolic acid (MPA) with potent immunosuppressive properties. Mycophenolate stops T-cell and B-cell proliferation through selective inhibition of the de novo pathway of purine biosynthesis. In vivo, the active metabolite, MPA, reversibly inhibits inosine 5'-monophosphate dehydrogenase, an enzyme involved in the de novo synthesis of guanine nucleotides. MPA displays high lymphocyte specificity and cytotoxicity due to

Lymphocyte Immune Globulin, Anti-thymocyte Globulin



Sources (1)

1 Path

5.00



UMLS Semantic Type: STY:T116; UMLS Semantic Type: STY:T121; UMLS Semantic Type: STY:T129

Prednisone



Publications (93)
Clinical Trials (5)
Sources (7)

18 Paths

5.00



It's important to track provenance

Direct Path 1.b Evidence

Mycophenolate Mofetil treats or applied or studied to treat Dysphagia · **Mark as Unseen**

Myophenolate Mofetil treats or applied or studied to treat Dysphagia phenotype of Immune-mediated...

Publications

Knowledge Sources ⓘ

Showing 1-2 of 2 Publications

Knowledge Level ⓘ

All

Curated

Text-Mined

TITLE	DATE(S) ▾	JOURNAL	SNIPPET	SOURCE & RATIONALE
High-dose Cyclophosphamide Without Stem Cell Rescue In Immune-mediated Necrotizing Myopathies.	2017	Neurology(r) Neuroimmunology & Neuroinflammation	To describe the experience managing treatment-refractory immune-mediated necrotizing myopathies (IMNM) with high-dose cyclophosphamide (HiCy) therapy. Fiv... Read More ↗	Text Mining Targeted Association API
Immunohistochemistry Of Affected Tissue May Guide Cgvhd Treatment Decisions.	2012	Bone Marrow Transplantation	Chronic graft-vs-host disease (cGVHD) myositis is a rare complication of hematopoietic SCT, for which the pathogenesis and optimal therapy are... Read More ↗	Text Mining Targeted Association API

Publications per Page

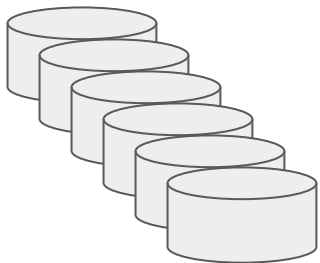
5 ▾

Previous

1

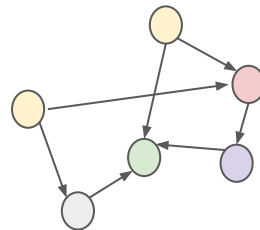
Next

Avoiding GIGO



ORION

ROBOKOP KG



ORION creates and exposes metadata

- Data Versions, of both input and output
- Tool Versions
- Processes
- Graph characteristics
- Evidence
- Provenance

Metadata exposed in machine readable formats

```
{
  "graph_id": "RobokopKG",
  "graph_name": "ROBOKOP KG",
  "graph_description": "ROBOKOP (KG) is an open-source biomedical KG that supports the ROBOKOP application and currently contains millions of biomedical relationships ..." ,
  "graph_url": "http://robokopkg.renci.org/browser/ ",
  "graph_version": "4901b2bc764444ea" ,
  "sources": [],
  "subgraphs": [],
  "build_status": "stable",
  "build_time": "07-30-24 23:37:31" ,
  "build_error": null,
  "final_node_count": 9800836,
  "final_edge_count": 137235468,
  "merged_nodes": 280904,
  "merged_edges": 4676433,
  "qc_results": {},
  "neo4j_dump":
  "https://stars.renci.org/var/plater/bl-4.2.1/RobokopKG/4901b2bc764444ea/graph_4901b2bc764444ea.db.dump  "
}
```

A dirty little secret

- Very few of the sources that ORION ingests have sufficient metadata.
- Almost none of them have metadata in a machine readable format.
- None of them have metadata in a machine readable format that conforms to any common standard.

Croissant: A Metadata Format for ML-Ready Datasets

Mubashara Akhtar⁷, Omar Benjelloun⁴, Costanza Conforti⁴, Pieter Gijsbers¹², Joan Giner-Miguel¹³, Nitisha Jain⁷, Michael Kuchnik⁸, Quentin Lhoest⁵, Pierre Marcenac⁴, Manil Maskey⁹, Peter Mattson⁴, Luis Oala³, Pierre Ruysen⁴, Rajat Shinde¹⁰, Elena Simperl^{7,11}, Geoffry Thomas^{4,6}, Slava Tykhonov², Joaquin Vanschoren¹², Jos van der Velde¹², Steffen Vogler¹, Carole-Jean Wu^{8*}

*Authors in alphabetical order

¹Bayer, ²DANS-KNAW, ³Dotphoton, ⁴Google, ⁵Hugging Face, ⁶Kaggle, ⁷King's College London, ⁸Meta, ⁹NASA,

¹⁰NASA IMPACT & UAH, ¹¹Open Data Institute, ¹²TUE & OpenML, ¹³Universitat Oberta de Catalunya

[paper](#)

```

1 { "@type": "sc:Dataset",
2   "name": "PASS",
3   "dct:conformsTo":
4     "http://mlcommons.org/croissant/1.0",
5   "description": "PASS is a large-scale image dataset
6     that does not include...",
7   "citeAs": "@Article{asano21pass, title = \"PASS: An
8     ImageNet replacement...\",
9   "license": "cc-by-4.0",
10  "url": "https://www.robots.ox.ac.uk/~vgg/data/pass/"}

```

Figure 3: Dataset metadata for the PASS dataset.

```

1 { "@id": "pass0",
2   "@type": "cr:FileObject",
3   "contentUrl":
4     "https://zenodo.org/6615455/PASS.0.tar",
5   "sha256": "0be3a104d6257d83296460b419f82c71",
6   "encodingFormat": "application/x-tar",
7   { "@id": "image-files",
8     "@type": "cr:FileSet",
9     "containedIn": { "@id": "pass0" }
10    "includes": "*.jpg",
11    "encodingFormat": "image/jpeg"},
12   { "@id": "metadata",
13     "@type": "cr:FileObject",
14     "contentUrl":
15       "https://zenodo.org/6615455/pass_metadata.csv",
16     "sha256": "0b033707ea49365a5ffdd14615825511",
17     "encodingFormat": "text/csv"}

```

Figure 4: Definitions of Resources for the PASS dataset.

```

1 {
2   "@id": "images",
3   "@type": "cr:RecordSet",
4   "key": "images/hash",
5   "field": [
6     {
7       "@id": "images/image_content",
8       "@type": "cr:Field",
9       "dataType": "sc:ImageObject",
10      "source": {
11        "fileSet": { "@id": "image-files" },
12        "extract": { "fileProperty": "content" }
13      }
14    },
15    {
16      "@id": "images/hash",
17      "@type": "cr:Field",
18      "dataType": "sc:Text",
19      "source": {
20        "fileSet": { "@id": "image-files" },
21        "extract": { "fileProperty": "filename" },
22        "transform": { "regex": "([^\s/]*).*\.jpg" }
23      },
24      "references": {
25        "fileObject": { "@id": "metadata" },
26        "column": "hash"
27      }
28    },
29    {
30      "@id": "images/coordinates",
31      "@type": "cr:Field",
32      "dataType": "sc:GeoCoordinates",
33      "subField": [
34        {
35          "@id": "images/coordinates/latitude",
36          "@type": "cr:Field",
37          "source": {
38            "fileObject": { "@id": "metadata" },
39            "column": "latitude"
40          }
41        },
42        {
43          "@id": "images/coordinates/longitude",
44          "@type": "cr:Field",
45          "source": {
46            "fileObject": { "@id": "metadata" },
47            "column": "longitude"
48          }
49        }
50      ]
51    }
52  ]
53 }

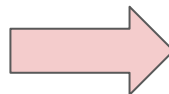
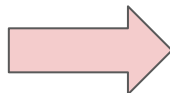
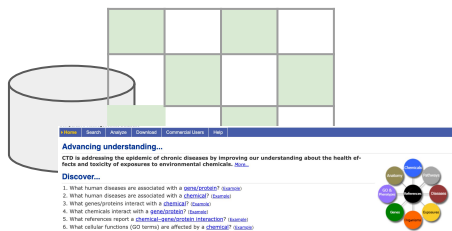
```

Figure 5: Definition of a RecordSet that joins images and structured metadata from the PASS dataset.

~~What is the role of Data Management in the age of AI?~~

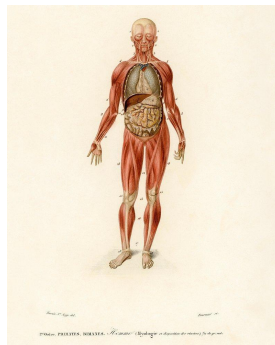
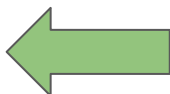
What is the role of AI in the age of Data Management?

We are developing a prototype “Auto-croissant” pipeline



What is wrong with my croissants? : r/AskBaking

Visit >



Thanks to

RENCI

Suparna Goswami
Yaphet Kebede
Yogesh Kale
Gaurav Vaidya
Evan Morris
James Chung
Kathleen Carter
Ola Olasunkanmi

UNC

Alex Tropsha
Marcello DeLuca
Kushal Koirala
Elvin Ding
Melissa Haendel
Niko Matentzoglu

Every Cure

David Fajgenbaum
Pascal Brokmeier
Lee Lancashire
Elliott Sharpe

Funding provided by

NHLBI (DugBot)
ARPA-H (MATRIX)