Data Knights Meetup 2024

Using Knowledge Graphs in Drug Repurposing

The Evolution from Traditional KGs to Foundation Models

Pascal Brokmeier

Who is Pascal

Work

- Head of Engineering at **Every Cure**
- Previously Principal Data Engineer at **Quantum Black**
- And of course … a **Data Minded** Engineer as well 😍

Education

- Studied in Information Systems at University of Cologne ("Duales Studium")

Life

- Born in Germany
- Live in the Netherlands
- Always keen to meet new people, **do come say hi** later!

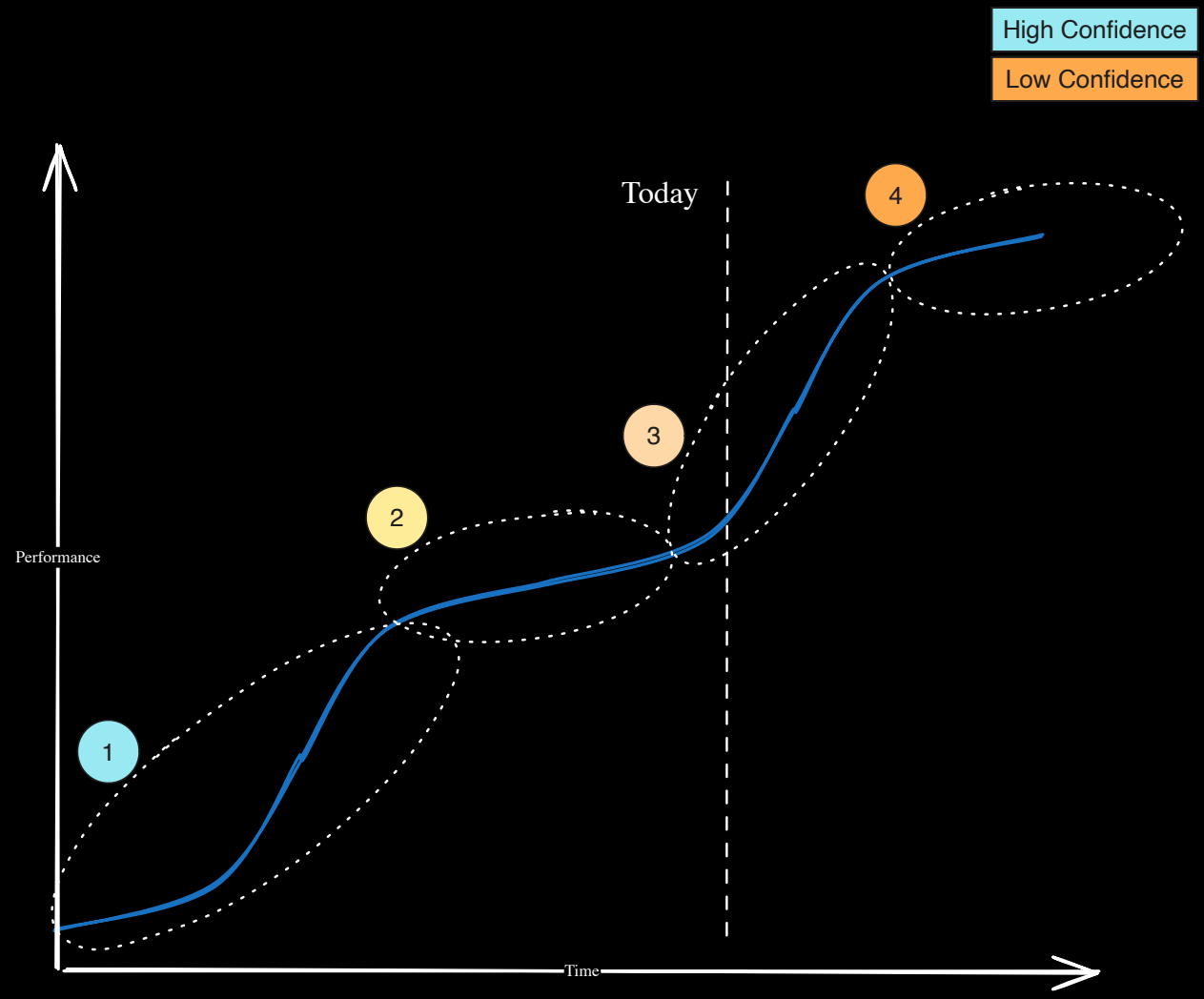
Our plot for today

First Act: The First S Curve - Traditional Knowledge Graphs and Machine Learning for Drug Repurposing

Second Act: Transition - The Need for a Paradigm Shift

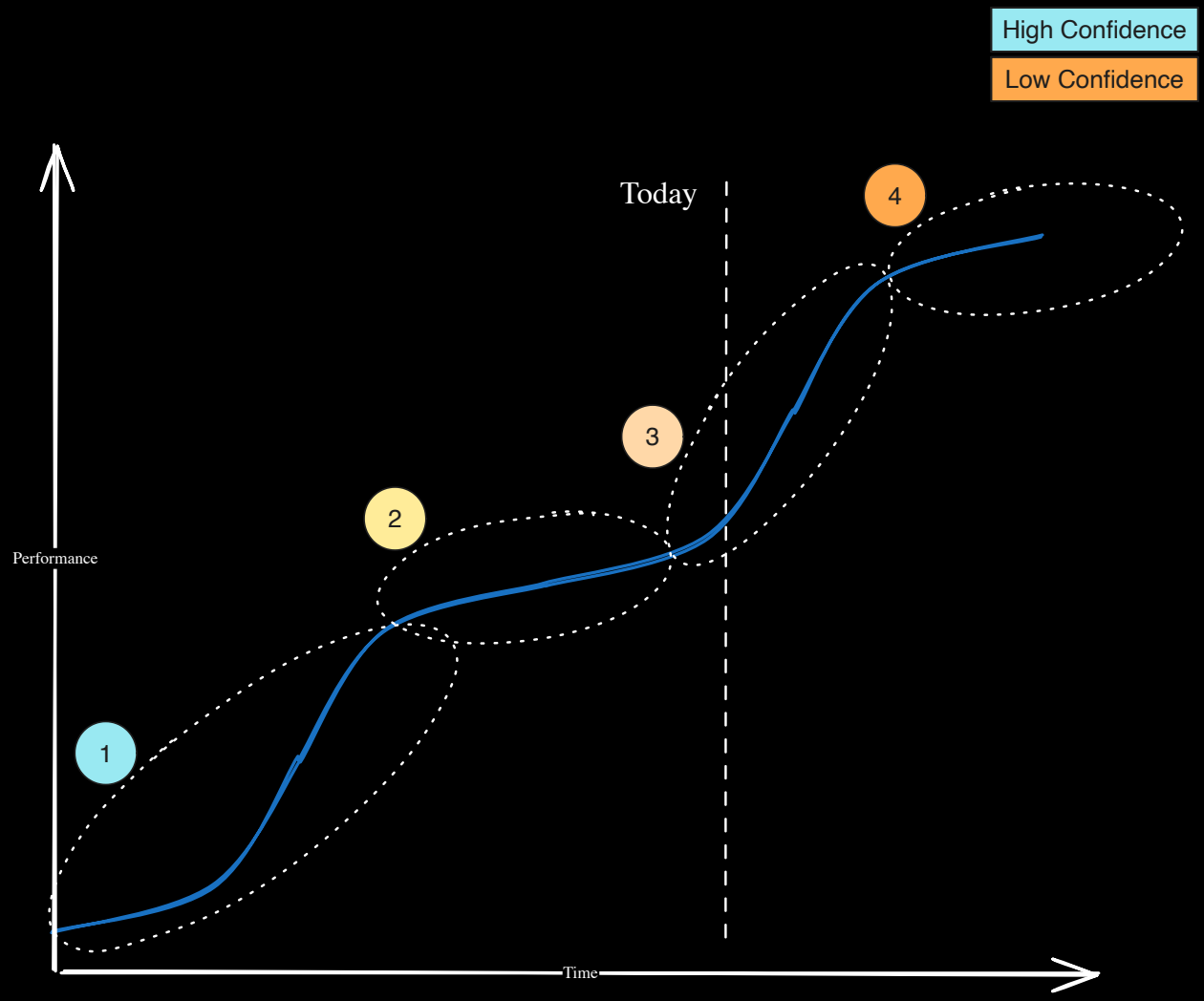
Third Act: Foundation Models stealing the show

Fourth Act: The place of Knowledge Graphs in the Future



Act 1: The First S Curve

Traditional Knowledge Graphs and Machine Learning for Drug Repurposing



1990+: Low Hanging Fruit are running out

Drug Discovery is becoming ever more difficult

- R&D costs soaring
- New drugs not keeping pace

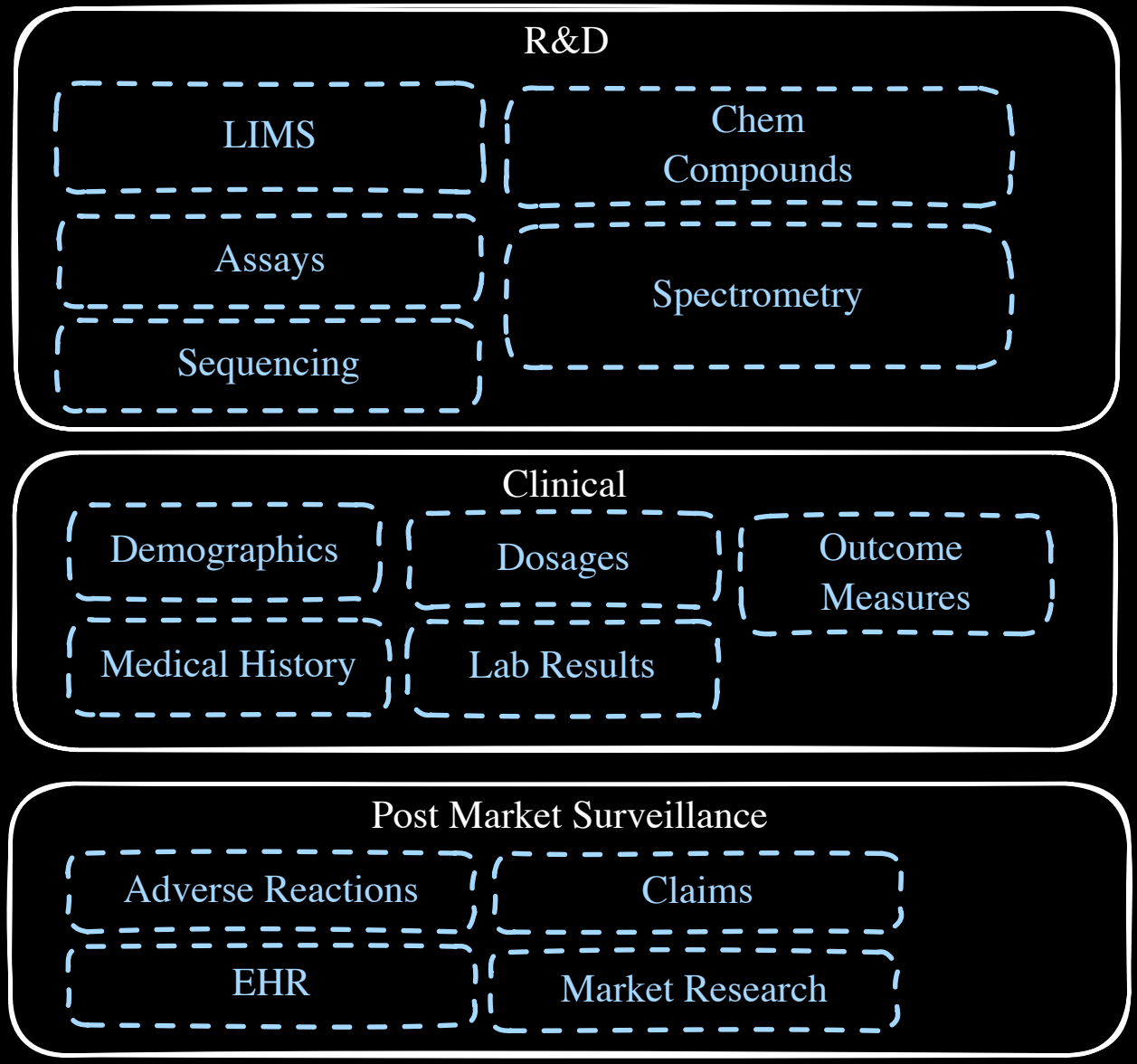
Same time: Data availability explosion

- Electronic records
- growing body of scientific literature
- Genomics & Imaging data
- Wearables

What if we could use this data 🤔

1990 - 2010: Silos of Knowledge being built up

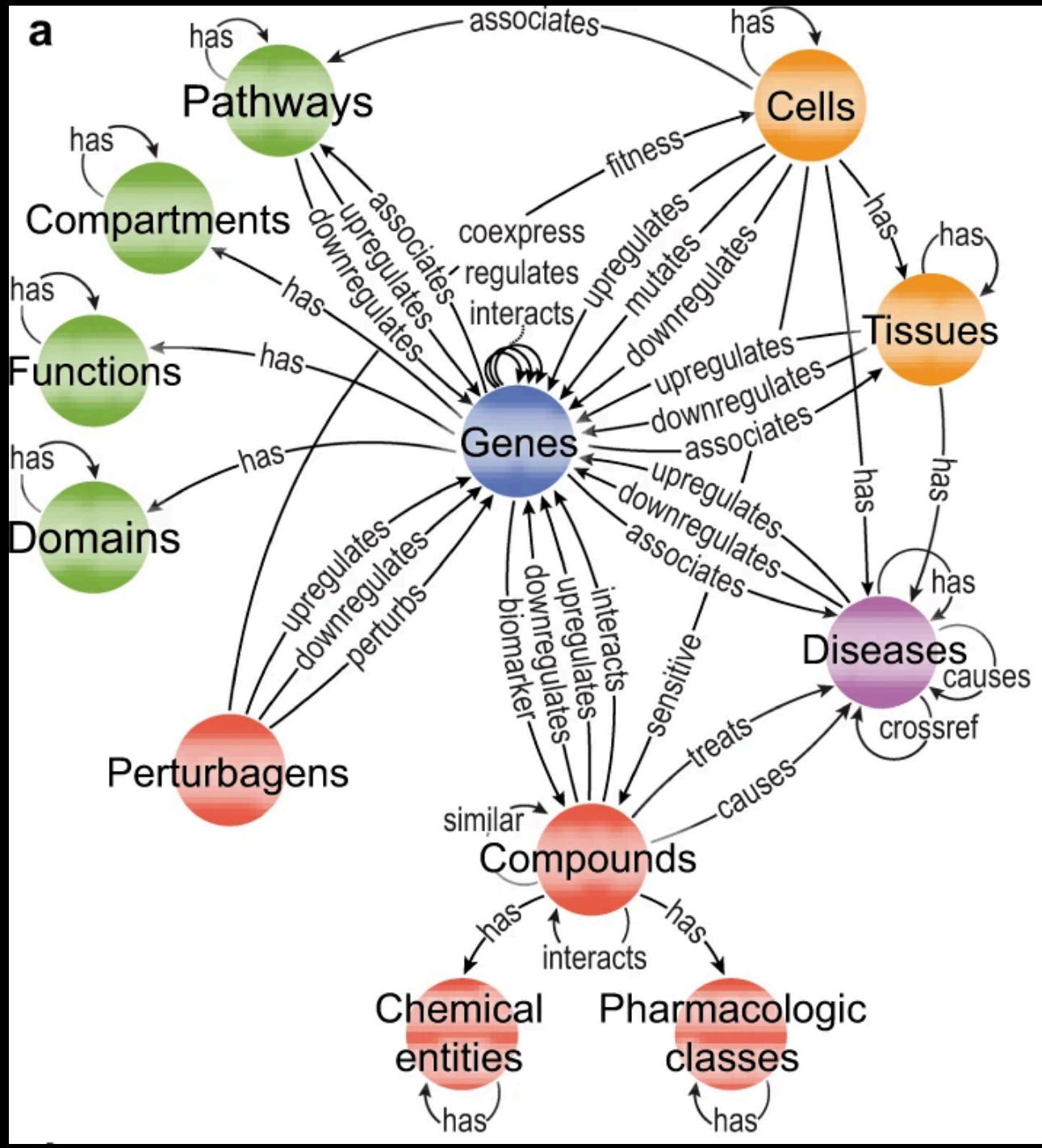
- Various domains building up their data domains
- Connectivity between them mostly missing
- Progress is made but unified view lacking



2010s: Traditional Knowledge Graphs: Early Wins

- Knowledge Graphs (KGs) as specialized databases unifying data across domains
- Connect entities: diseases, genes, proteins, drugs
- Early examples:
  - Hetionet: Integration of multiple biomedical databases
  - OpenBioLink: Integration of multiple biomedical databases
  - RTX: Integration of multiple biomedical databases

Source: https://www.nature.com/articles/s41467-022-33026-0



Arrival of Specialized Knowledge Graphs

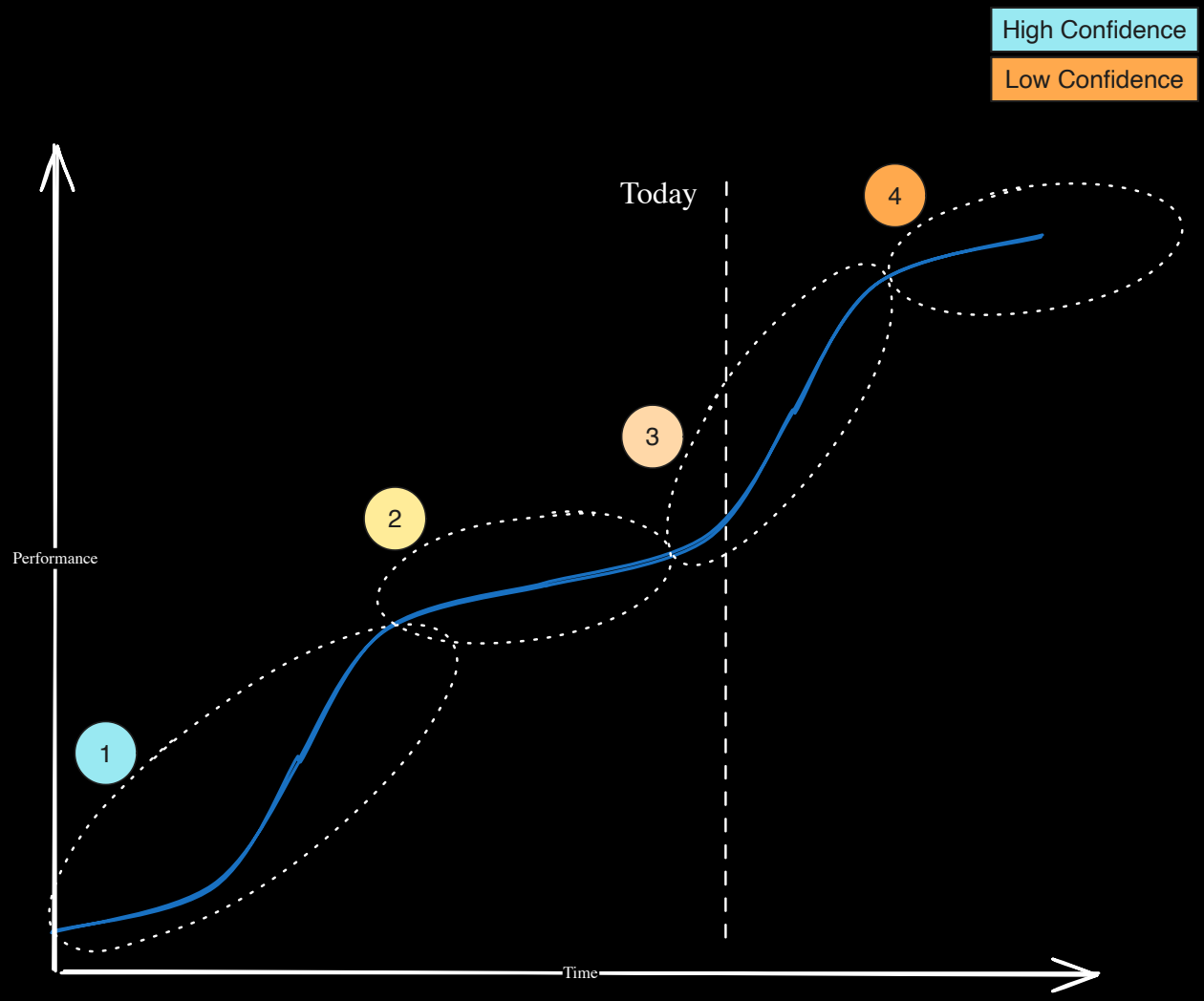
- SPOKE: Focusing on high curation quality
- GARD: Designed for rare diseases
- RTX KG2: Prioritizing number of sources and data categories ingested
- PRIME: Leveraging Embedding distances for clique merging
- ...

There is a lot of work being done but somehow this still feels like a *breadth first search*

We are not really getting closer to actual patient impact

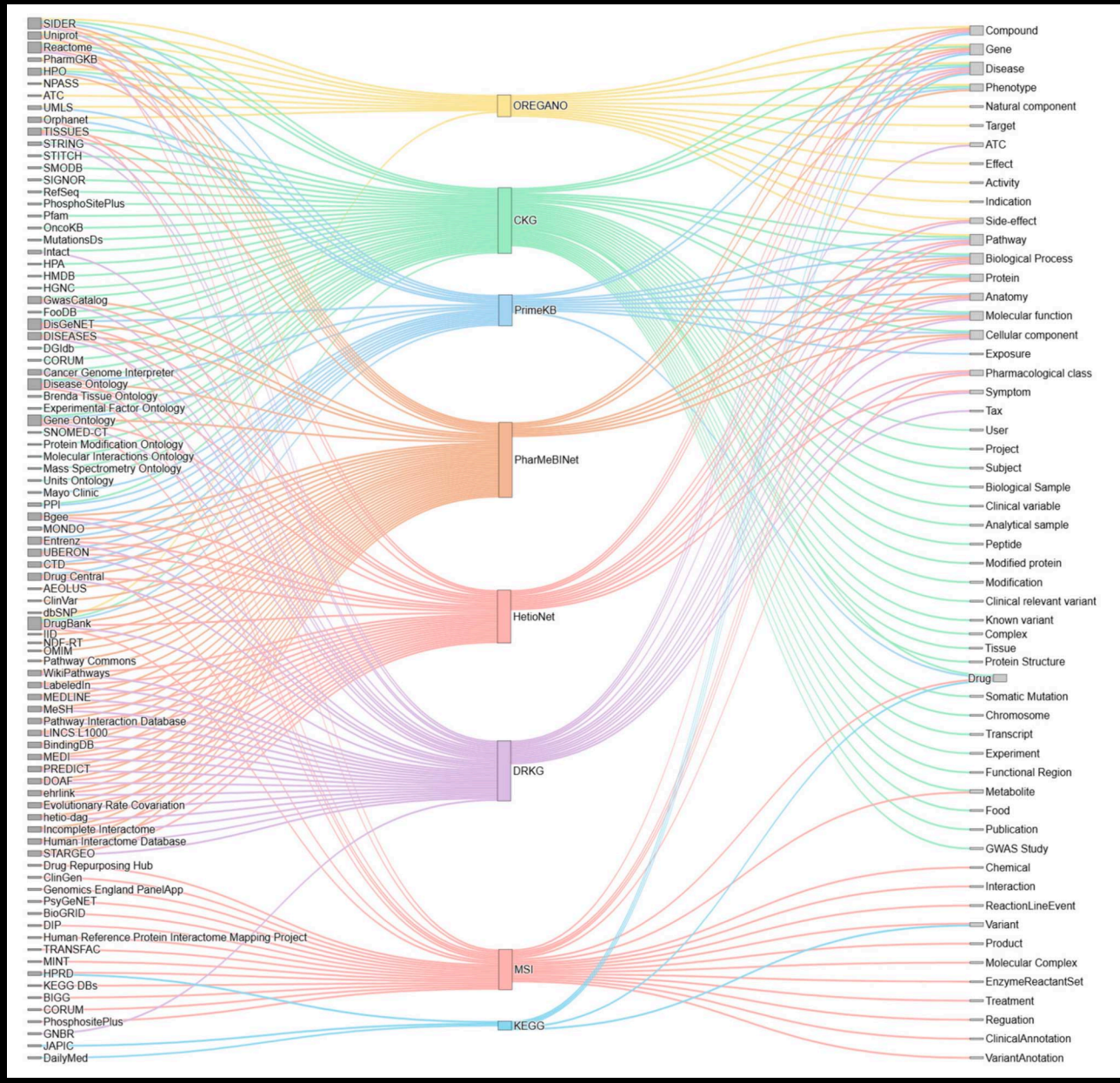
Act 2: The Transition

KGs provide a great scaffold but their creation is still painfully manual



Problem 1: Everyone copies from the same places

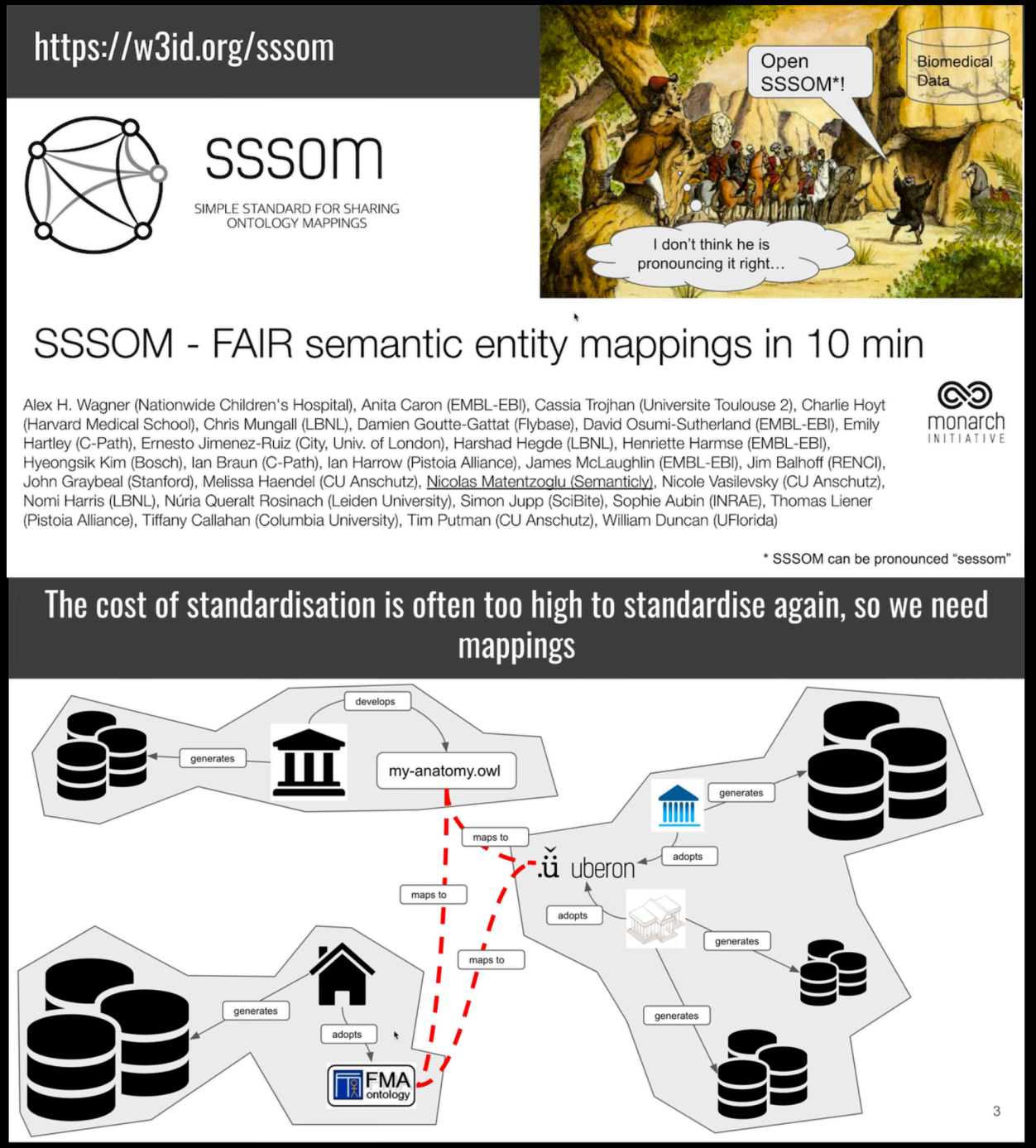
- KGs need to ingest existing data, often from the same sources
- However, how data is ingested differs greatly
- So if more is not always better and ingest does not = ingest
- How about unifying standards?



Problem 2: KGs are distinct so we need IDs

But IDs require someone to hand them out

- We originate from a siloed world
- each domain already had its own ID systems
- now also KGs have their own ontology systems



Flux Dev Generated

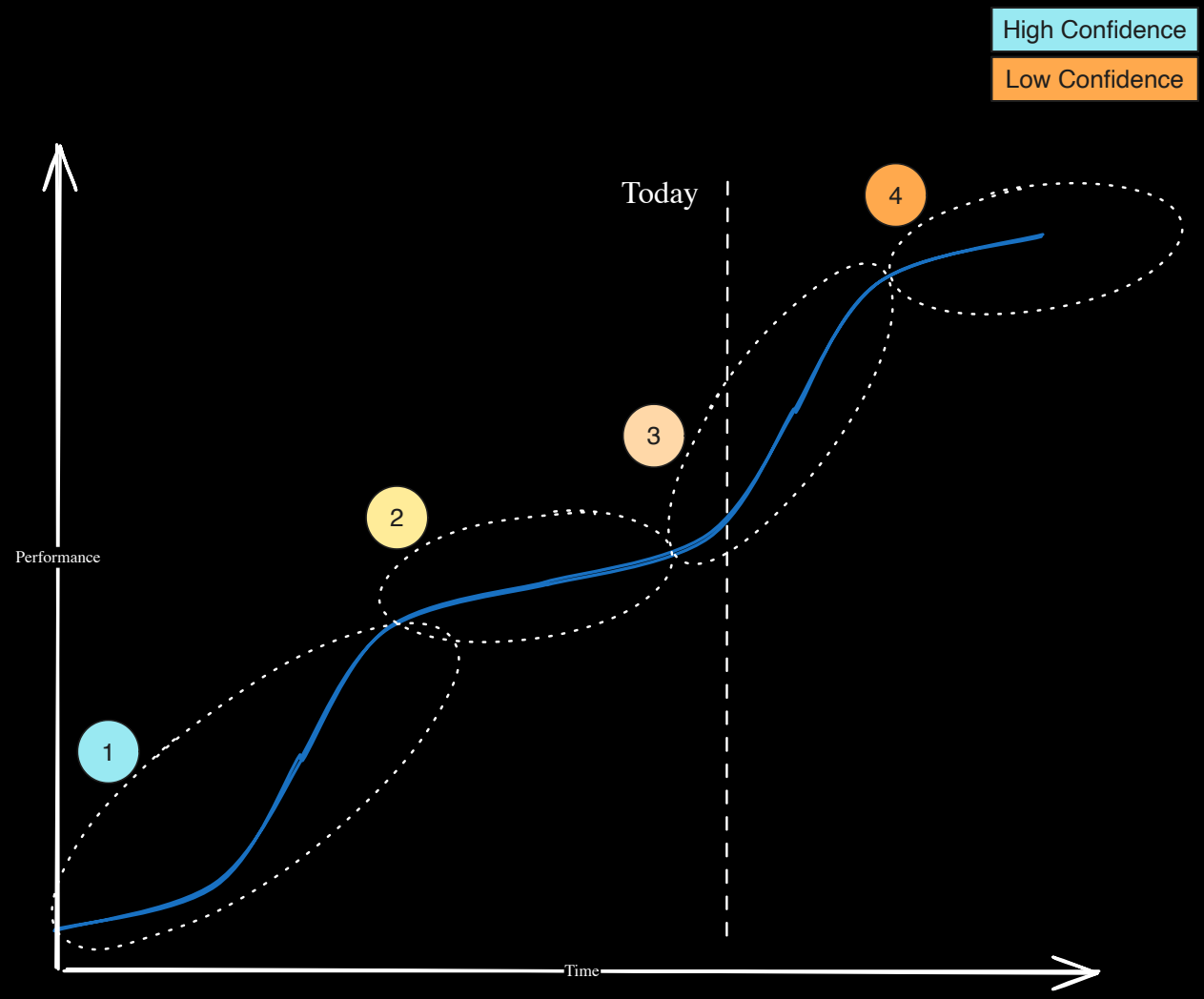
Problem 3: KGs struggle encoding continous data

- KGs excel at encoding relationships and networks
- But a living organism is a messy continuous space
- Not easy to encode without losing too much information

Source: https://w3id.org/sssom

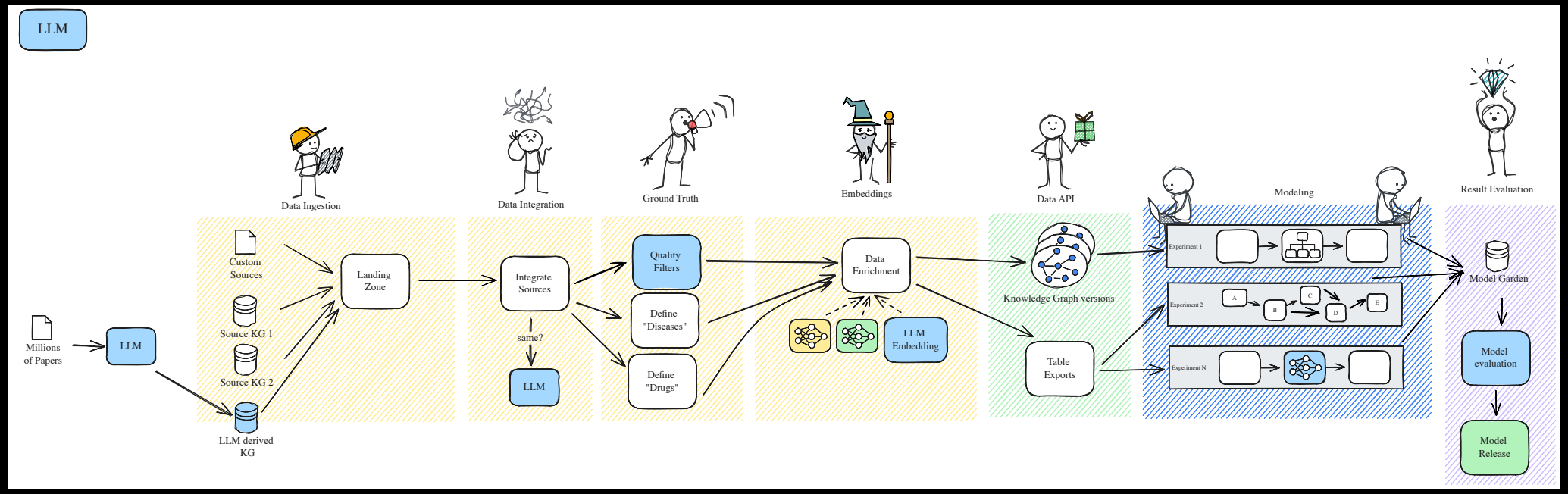
Act 3: The Rise of Foundation Models

Because really all we need is...




CHATGPT, DUDE.

Language models are being proposed at every step of the process

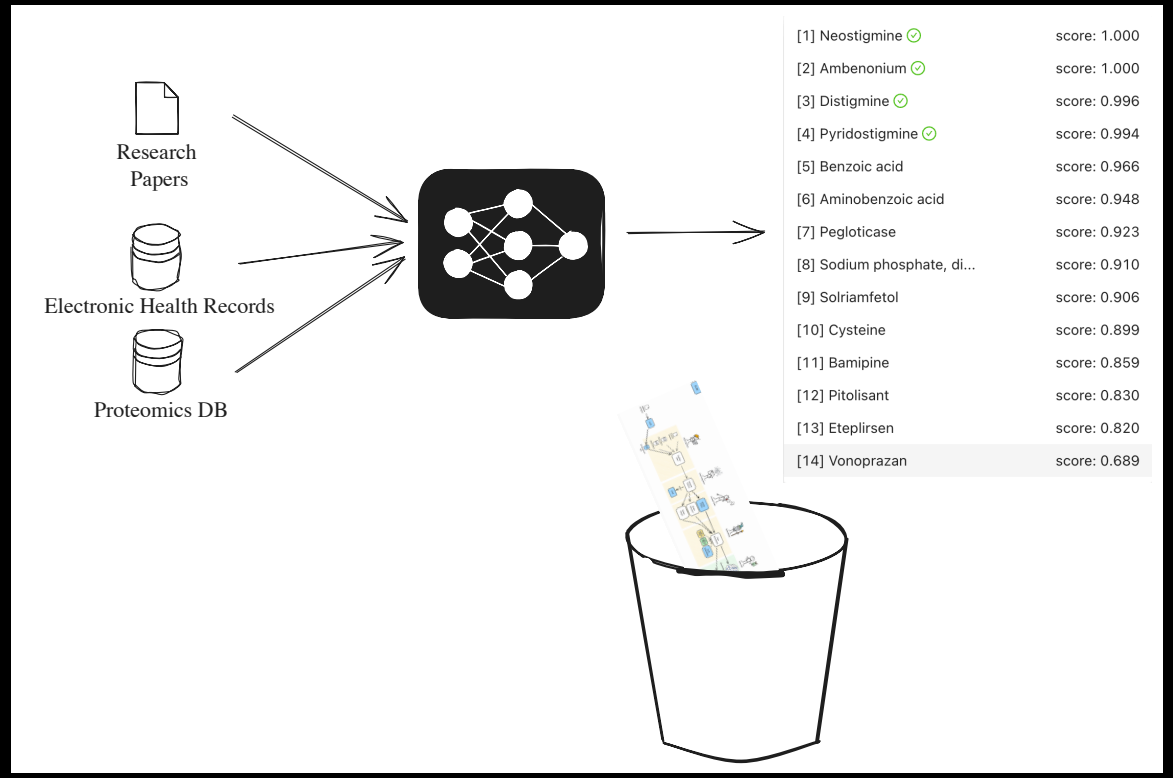


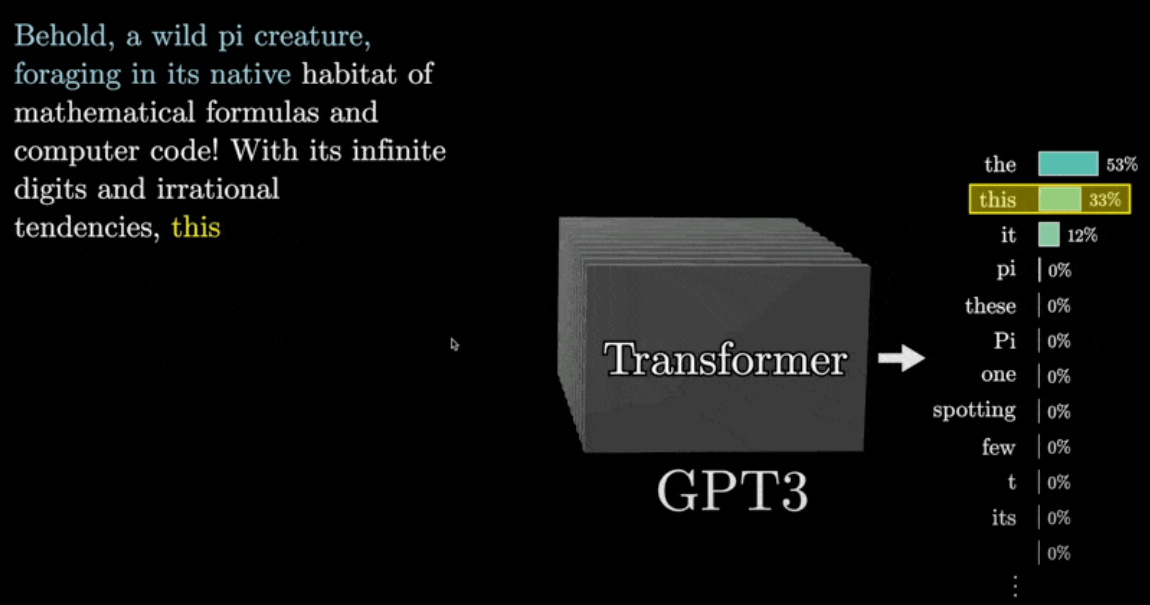
Initial successes leveraging LLMs but the KG representation of information remains

- Merging equivalent nodes: PRIME KG uses BERT based cosine similarity to merge nodes
- Verifying drug-disease predictions: LLMs have helped verify predictions by collecting evidence from text
- Generate new KGs: LLMs have been used to extract triples from scientific literature
- GNN for drug-disease predictions

However, we are still using the KG as an intermediate information representation

(Raw Take): Maybe *Attention is all you need*? Transformers may just directly generate the prioritized drug list for each disease





Independent of the approach, scalability, bias, explainability and verifiability remain hard problems

- Scalability issues with growing biomedical data when KG representation remains
- Data bias affecting predictions
- Lack of explainability ("black box" problem)
- Limited integration of real-world data

Most of all: **Generating predictions is easy, verifying if they are correct is hard**

Conclusions:

1. KGs broke the silos. Good!
2. But we are hitting their limitations: Manual curation, thinking in IDs, missing continuous data
3. Foundation Models might be a total replacement, rather than just a Turbo Charger we tack onto it
4. But unless you solve the verification problem all you have is a black box

Some useful links

- TXGNN
- Monarch initiative, doing the integation heavy lifting
- Open Evidence, medical RAG, done well
- Biolink Model defining a KG language
- OpenTargets, mapping drug targets to diseases

Last but not least: Keep an eye out for our GitHub organisation, we will open source our work soon

Link to this presentation

Thank You!

Questions?