Accionlabs

Accion INNOVATION SUMMIT 2023

02-05 March 2023,
Sofitel Dubai
The Palm Jumeirah
Dubai

Accion
INNOVATION
SUMMIT 2023

Accionlabs

Knowledge Graphs on Steroids, with Natural Language Queries

Accion
INNOVATION
SUMMIT 2023

Accionlabs

Dr Prof. Bhushan Bonde

Innovation Dev Head
Prof. Comp Biology (AI/ML and Quantum Computing)
Scientific Data Officer

Dr. Bhushan Bonde leads the Innovation Development for Pharmaceutical R&D. He received Ph.D. in Systems Biology and Mathematical Modelling, with an interdisciplinary background in Computational biology, Biotechnology and Mathematical Biology from Oxford (2006).

He also had an interdisciplinary M. Tech in Bioprocess technology/Chemical Eng. (2001) and Bachelor of Pharmacy (1998) and had been working across healthcare and R&D Pharma-IT Industry, University and Government Research Institutes with 20 years of experience in Computational Biology.

Table of Contents

Introduction to Knowledge Graphs

Advantages of KG

Use cases

KG Biological network

ChatGPT

Summary

Accion INNOVATION SUMMIT 2023

Accionlabs

Knowledge Graphs

Knowledge Graphs and Data Mesh : Introduction

Accionlabs

Knowledge graphs are special graph with entities:

- objects, events, or concepts that are interlinked by certain knowledge or relations between the entities.

Knowledge Graphs (KG) enable

- efficient means of **data management**,
- **Compact storage, and retrieval** with almost no data duplication,
- provide a single snapshot of the data held by the organisation.
- The added benefit of Graphs - **efficient, impactful visualisation** of the data

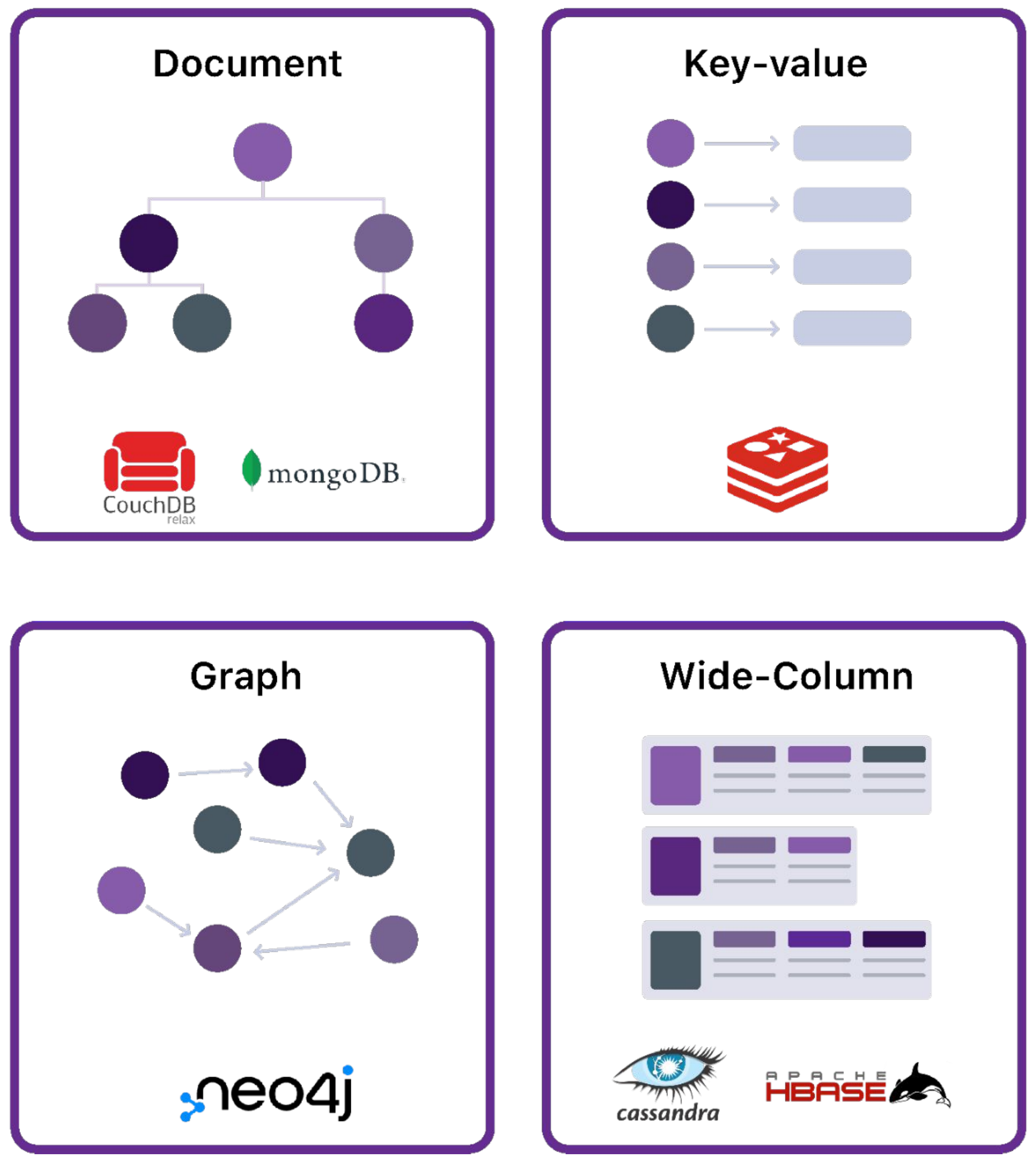
Pattern queries : Cypher

```
MATCH (p:Person)-[:LIVES_IN]->(c:City), (p:Person)-[:NATIONAL_OF]->(EUCountry)
RETURN p.first_name, p.last_name, c.name, c.state
```

Graph databases vs Relational and NoSQL databases

Accionlabs

Non-relational (NoSQL)



Azure Cosmos DB

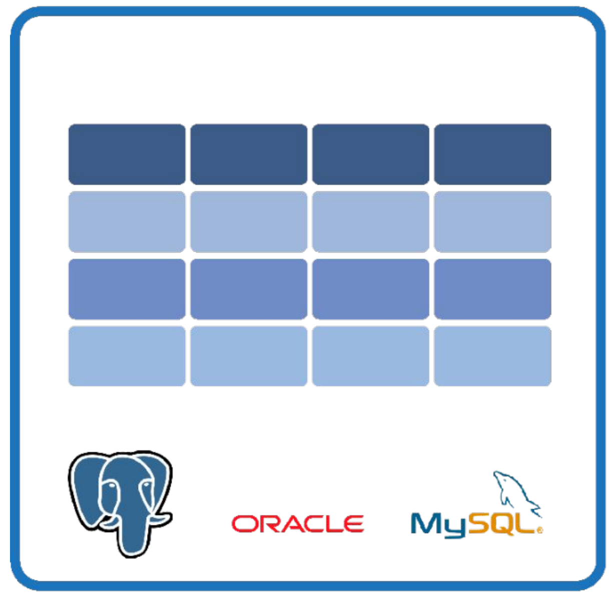
Amazon Neptune

Dgraph

TigerGraph CLOUD

OrientDB

Relational (SQL)



	Graph Mesh	Relational Databases	NoSQL Databases
Data Storage	Graph Storage Structure	Fixed, predefined tables with rows and columns	Minimal connected data support at the DB level
Data Modeling	Flexible Data Model/No schema	Schema model to be developed from a logical model (Data Normalisation)	No schema but may not be for enterprise architectures/security
Query Performance	High performance regardless of number and depth of connections	Data processing speed slows with number of (inner/outer) joins	Performant but Relationships must be created at the application level
Query Language	**Cypher**: native graph query language (GraphQL) /SparQL	**SQL**: complexity grows as the number of joins increases	Different query languages; None is tailored to express relationships
Transaction Support	Retains ACID transactions	ACID transaction support	BASE transactions prove unreliable for data relationships
Processing At Scale	Inherently or highly scalable for pattern-based queries	Scales through replication, but it's slow, costly	Scalable, but data integrity isn't trustworthy

Advantages of KGs

Knowledge graphs have distinct advantages over the rational and other no-sql technologies:

- It combines **siloed data** sources, both from **structured and unstructured** databases.
- It provides **summary and insights, community or clusters** which are very important for decision making.
- KGs also convert tables to networks : **reveals true data pattern**
- KG represents the **holistic visualisation of the depth**, **breadth and diversity** knowledge of data by leveraging

  - data integration,
  - advanced analytics (from clustering to shortest path and newer link prediction)
  - deep learning with state of art **Graph Neural Network** (GNN) approaches

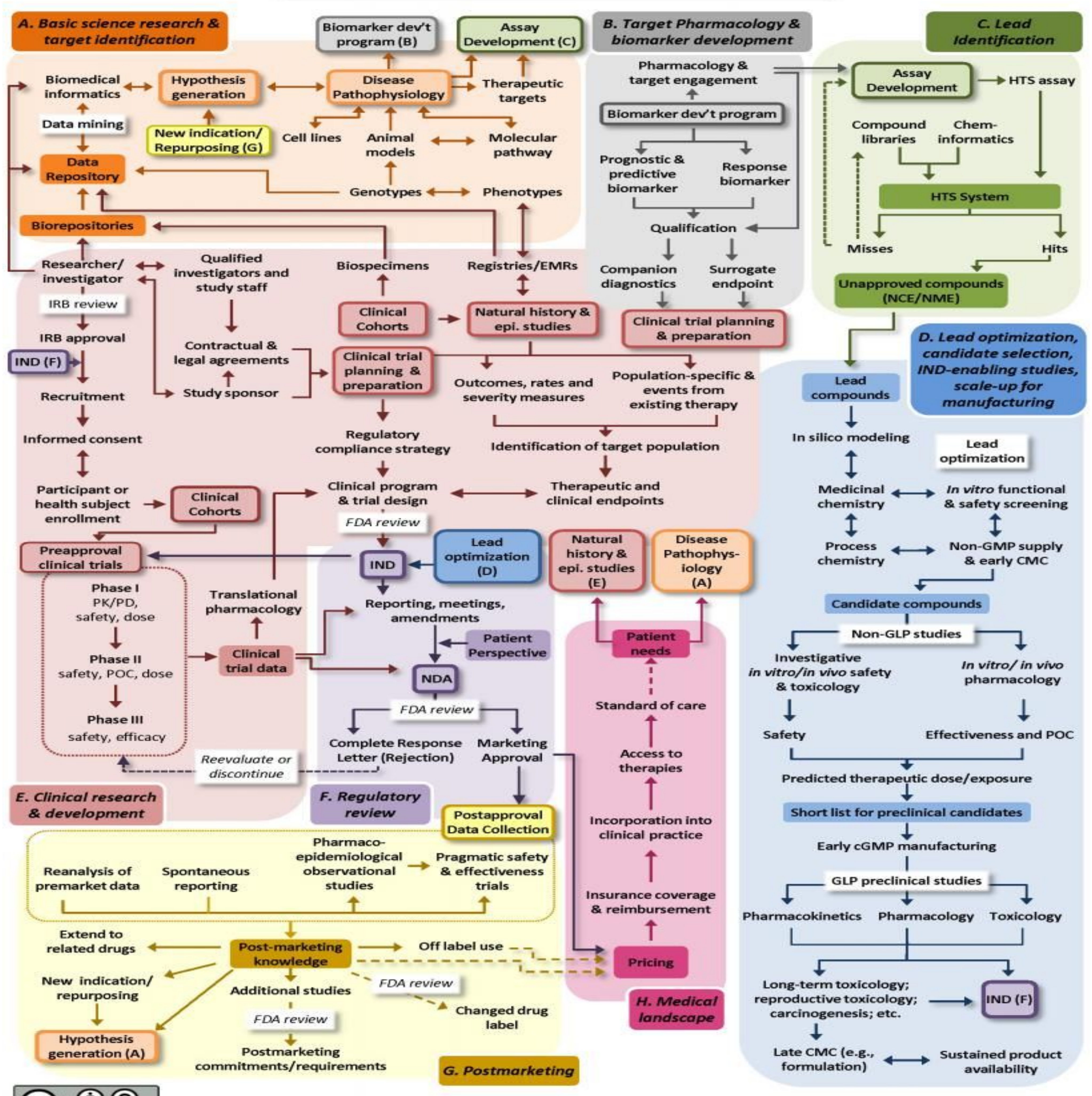
Accion INNOVATION SUMMIT 2023

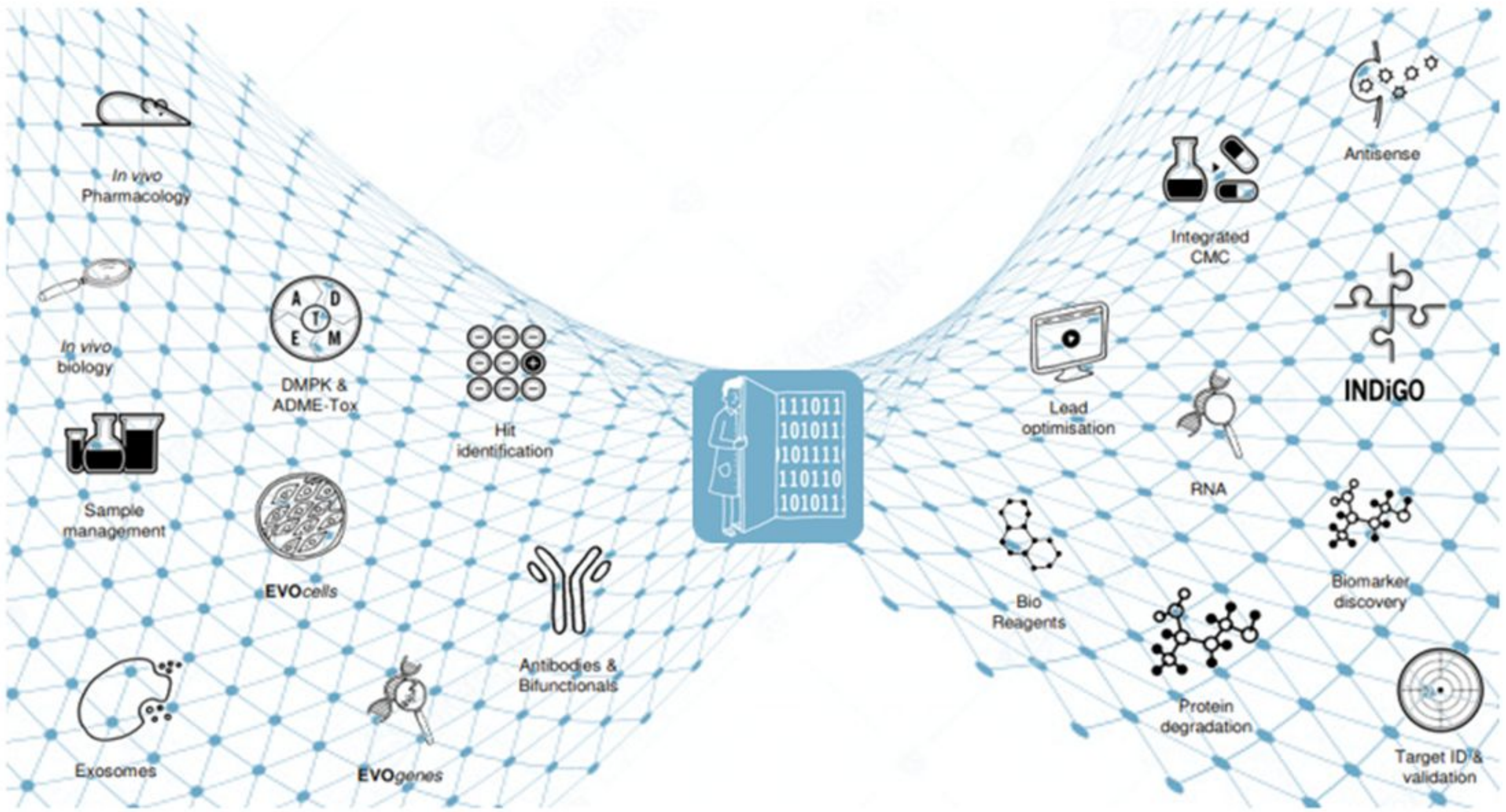
Accionlabs

Use cases

Whats wrong with New Drug Discovery Research ?

Disease Biology → Target Selection → Target Validation → Lead Discovery → Lead Optimization → Pre Clinical Developm ent → Clinical Developm ent → Regulatory Approval

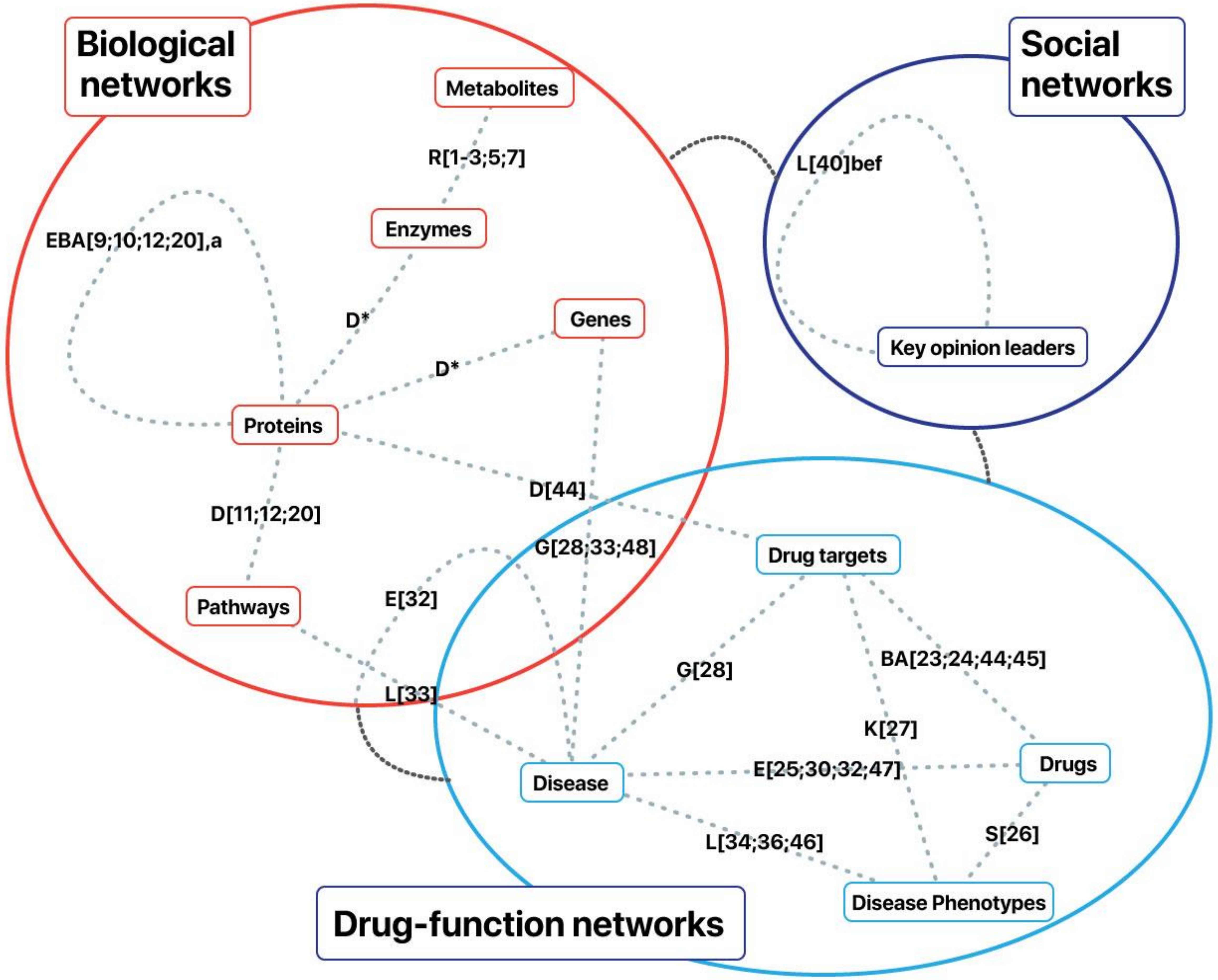




- Overall Timescale : 12 to 15 years / drug
- False discovery rate: 95% (early phase)
- Failure in efficacy: 30 % (early phase)
- Failure due to safely: 33 % (late Phase)
- Cost burdon : ~ 2 to 3 billion

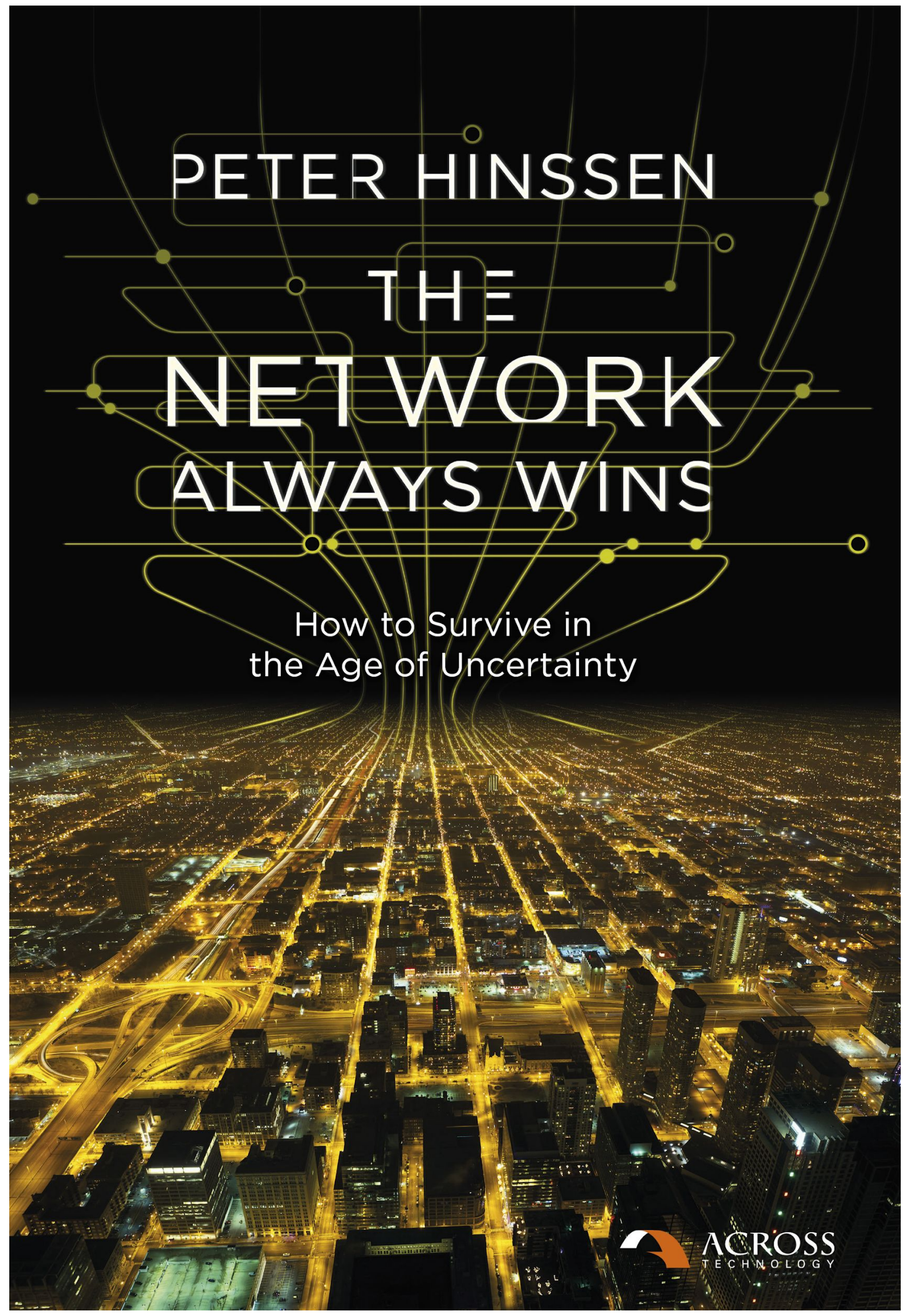
Knowledge Graph : Biological Entities

Accionlabs



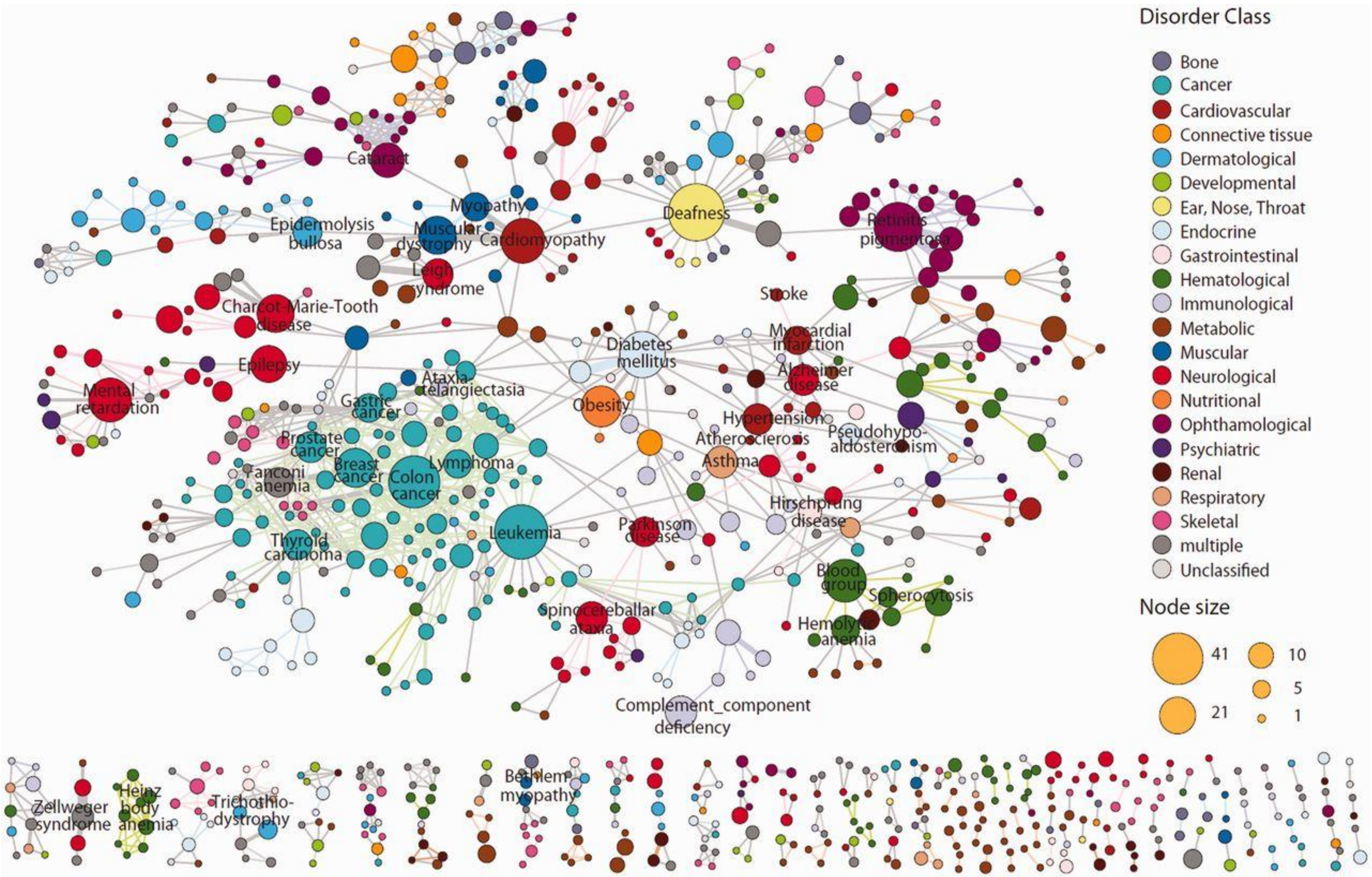
Biological entities ('nodes') of interest for drug discovery with their associations as Relationship ('edges')

A – molecular activity;
B – physical binding;
D – database repository;
E – gene expression;
G – genetic association;
K – knockout phenotype;
L – literature-based including ontologies and co-citations;
R – metabolic reaction;
S – side effect.



Ref: Hasan, Bonde et.al Drug Discovery Today (2012) Volume 17, Numbers 15/16, 869-874

KG: Disease landscape for Better Health and Drug Discovery



- Barabási lab (~2007) reported network-based disease analysis provides novel insights
- Most diseases share a common target (power law) : Common disease are well connected
- Neglected disease (bottom row) do not share common target or not well studied yet
- This was the first systematic effort in graph mining Gene-Disease network data

Limitation:

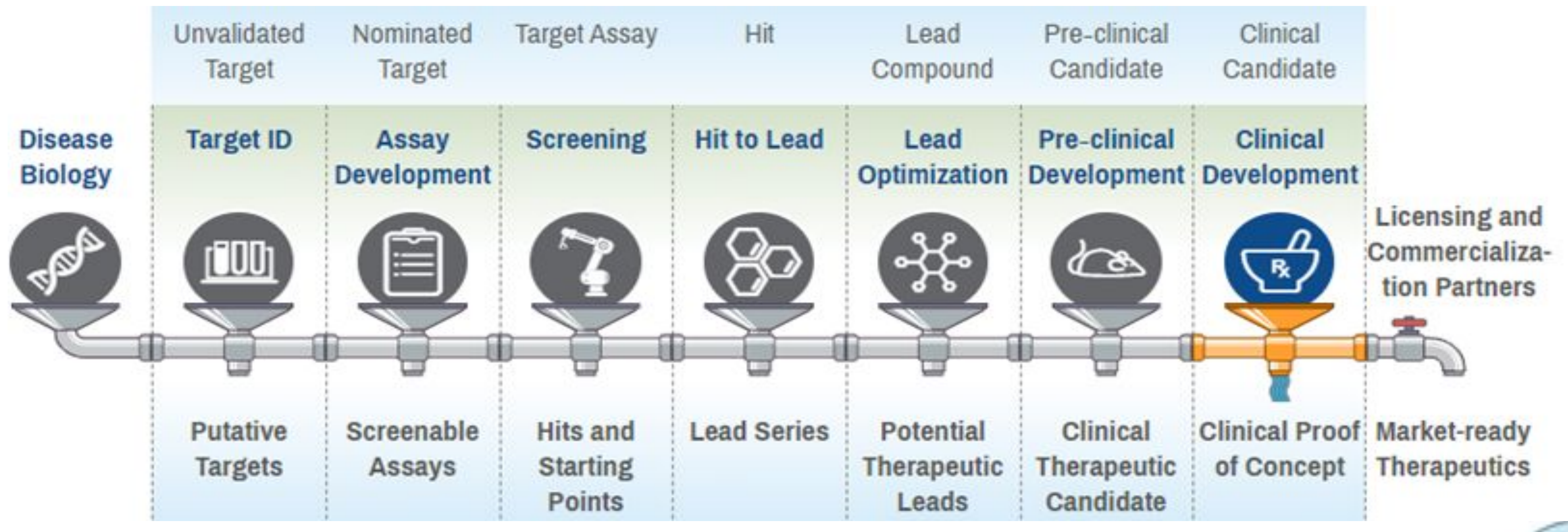
- Despite its novelties, its only qualitative in prediction capabilities
- No direct use/application possible in drug discovery
- Limited only to Gene and Diseases

Ref : Goh et al. The human disease network, Proc Natl Acad. Sci USA, 2007, 104, 8685-90

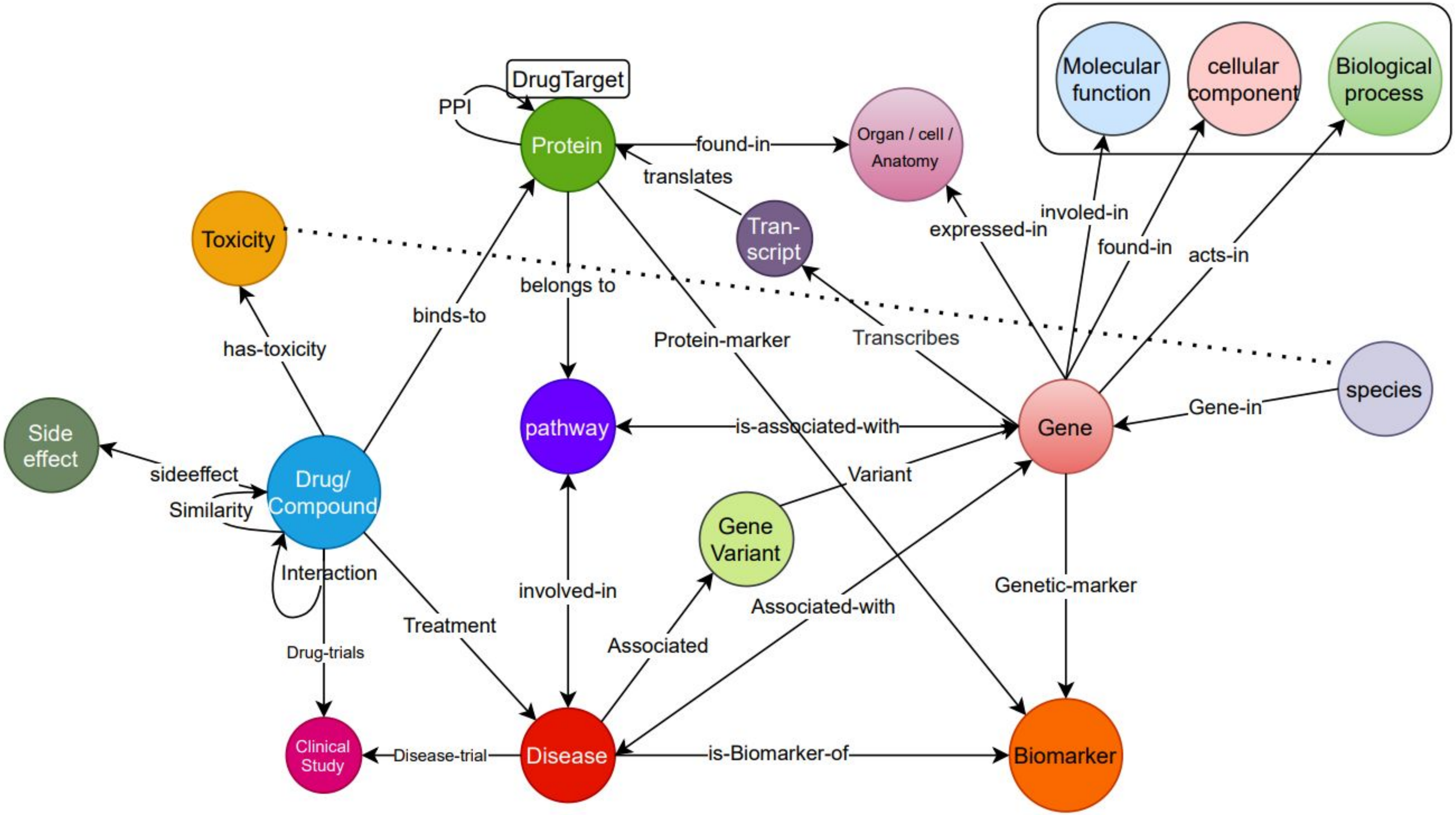
KG: (Biological) Data Lakes to Data Mesh

Accionlabs

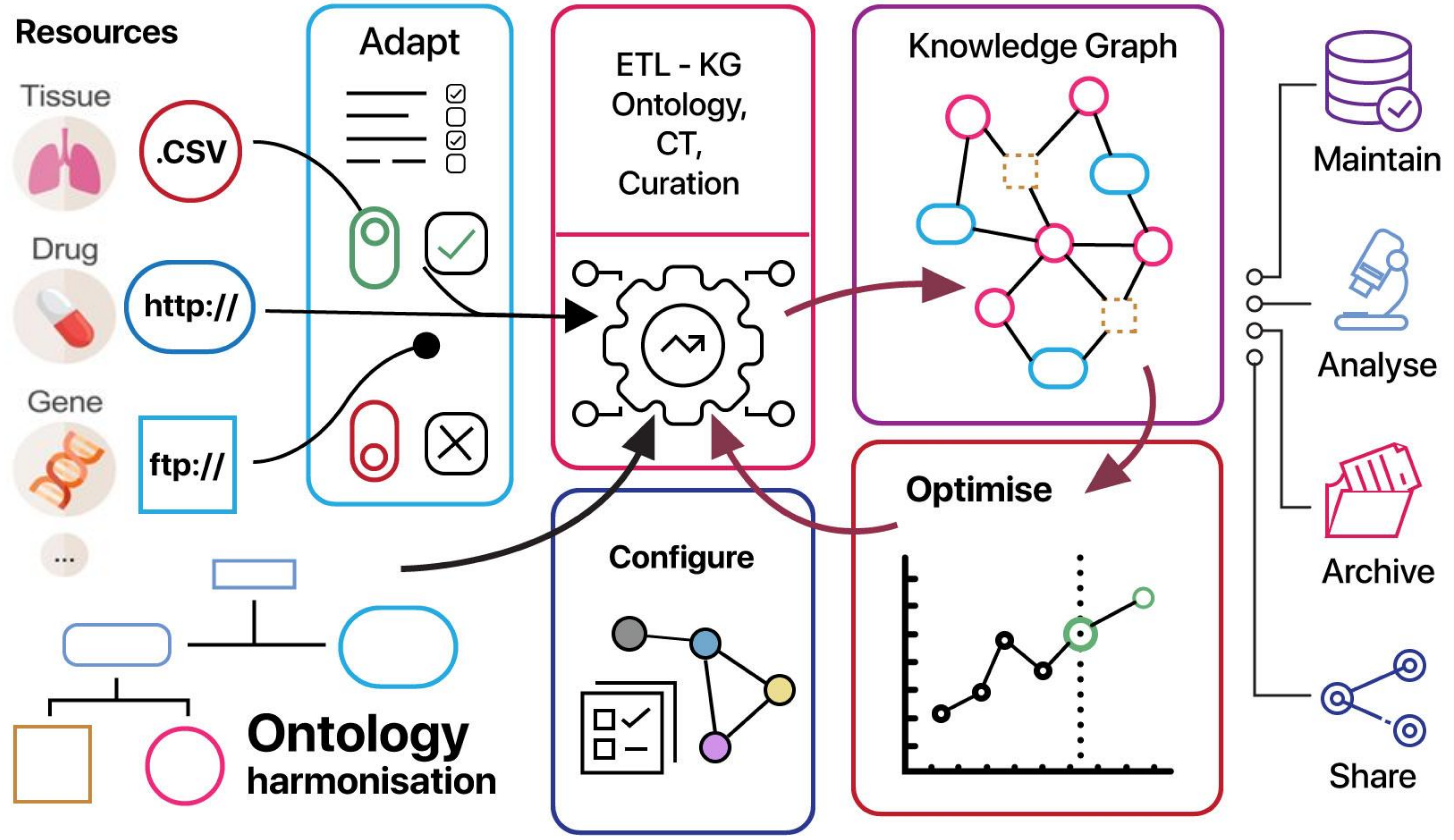
Disease Biology → Target Selection → Target Validation → Lead Discovery → Lead Optimization → Pre Clinical Development → Clinical Development → Regulatory Approval



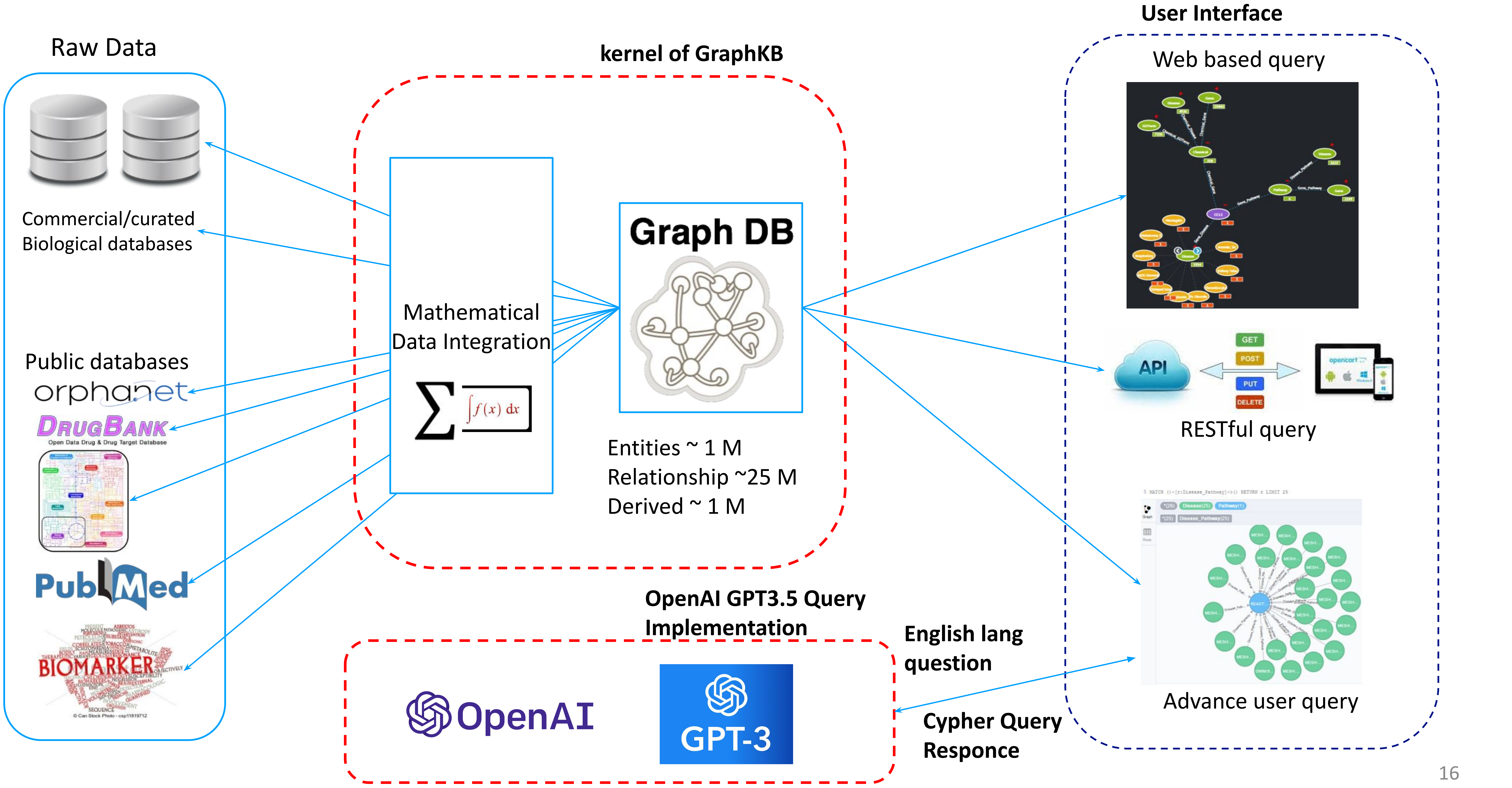
Bio Multiverse : Knowledge Graph for drug repositioning



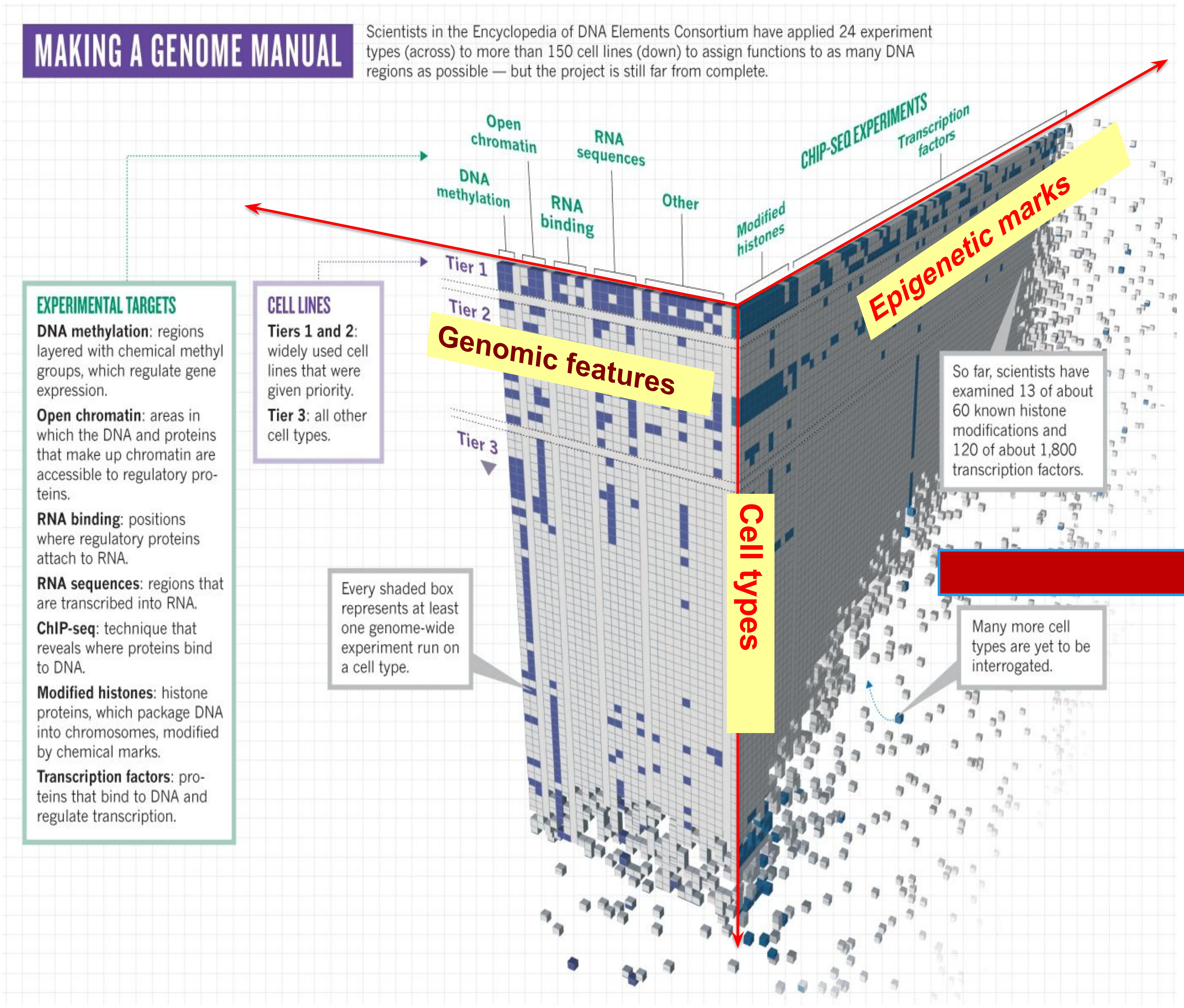
Building the Knowledge Graph : Life-cycle



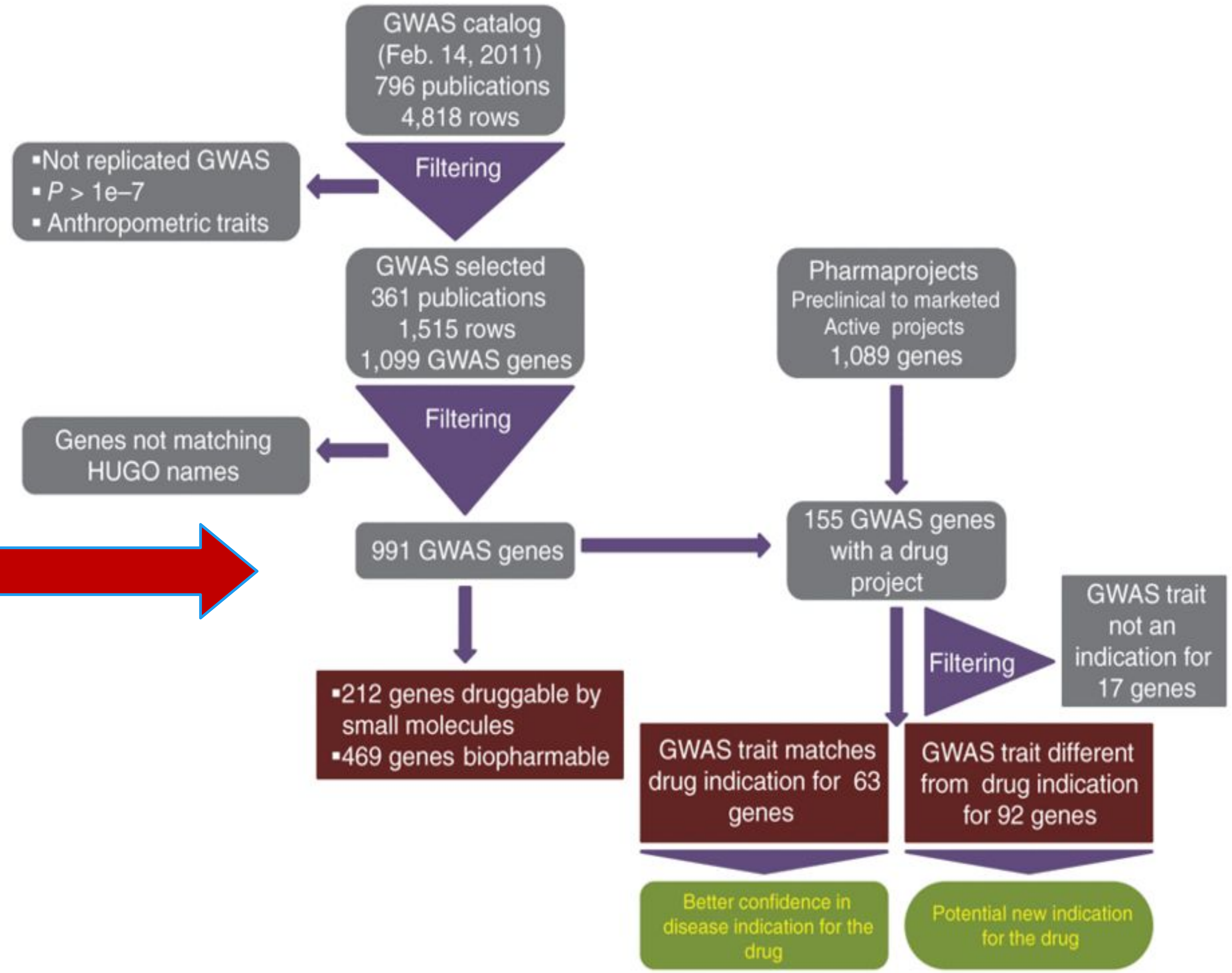
Knowledge Graph – Full Stack



Impact of Knowledge Graph and Data Mesh



Epigenetics landscape



Use of genome-wide association studies for drug repositioning

- Ref : Sanseau et. al., Nature Biotechnology 30, 317–320 (2012)

Accion INNOVATION SUMMIT 2023

Accionlabs

Large Language Models (LLM)

KG: Disease landscape for Better Health and Drug Discovery

Accionlabs

Which anatomies express migraine associated genes?

```
MATCH path =
(n0:Disease)-[:ASSOCIATES]-(n1)-[:PARTICIPATES]-(n2:BiologicalProcess)
WHERE n0.name = 'migraine'
WITH
[ size((n0)-[:ASSOCIATES]-()),
  size(()-[:ASSOCIATES]-(n1)),
  size((n1)-[:PARTICIPATES]-()),
  size(()-[:PARTICIPATES]-(n2))
] AS degrees, path, n2
WITH
   count(path) AS PC,
RETURN
  anatomy_id, anatomy_name, PC
LIMIT 5
```

anatomy_id	anatomy_name	PC
UBERON:0001645	trigeminal nerve	7
UBERON:0001785	cranial nerve	10
UBERON:0002363	dura mater	2
UBERON:0002925	trigeminal nucleus	4
UBERON:0002360	meninx	1

KG: Disease landscape for Better Health and Drug Discovery

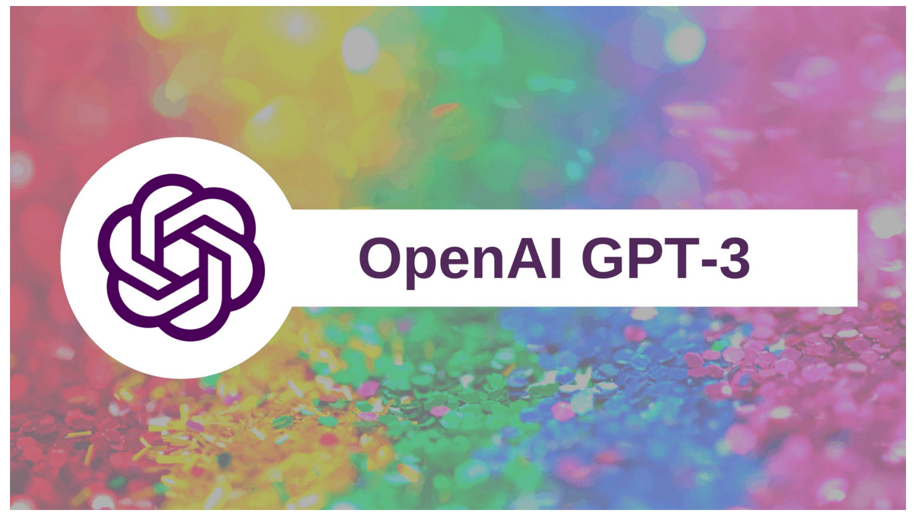
Accionlabs

Q1	Give me all oxidoreductase inhibitors active <100 nM in human and mouse
Q2	For a given compound, what is its predicted secondary pharmacology?
Q3	Given a target find me all actives against that target, and find and/or predict the polypharmacology of actives
Q4	For a given interaction profile, give me similar compounds
Q5	For molecules that contain substructure X, retrieve all bioactivity data in serine protease assays
Q6	For a specific target family, retrieve all compounds in specific assays
Q7	For a target, give me all active compounds with the relevant assay data
Q8	Identify all known protein–protein interaction inhibitors
Q9	For a given compound, give me the interaction profile with targets
Q10	For a given compound, summarize all similar compounds and their activities
Q11	Retrieve all data for a given list of compounds depicted by their chemical structure (SMILES) with options to match stereochemistry
Q12	For a given compound, which of its targets have been patented in the context of a disease?
Q13	For disease X, which targets have ligands in different stages of the development process with publications and/or patents describing these compounds?
Q14	Target druggability: compounds directed against target X have what indications?
	Which new targets have appeared recently in the patent literature for a disease?
Q15	Which chemical series have been shown to be active against target X?
	Which new targets have been associated with disease Y?
	Which companies are working on target X or disease Y?
Q16	Targets in Parkinson's disease or Alzheimer's disease are activated by which compounds?
Q17	For my specific target, which active compounds have been reported in the literature?
Q18	For pathway X, find compounds that agonize targets assayed in only functional assays with potency <1 MM
Q19	For the targets in a given pathway, retrieve the compounds that are active with more than one target
Q20	For a given disease, retrieve all targets in the pathway and all active compounds hitting those targets

Which model you prefer ?

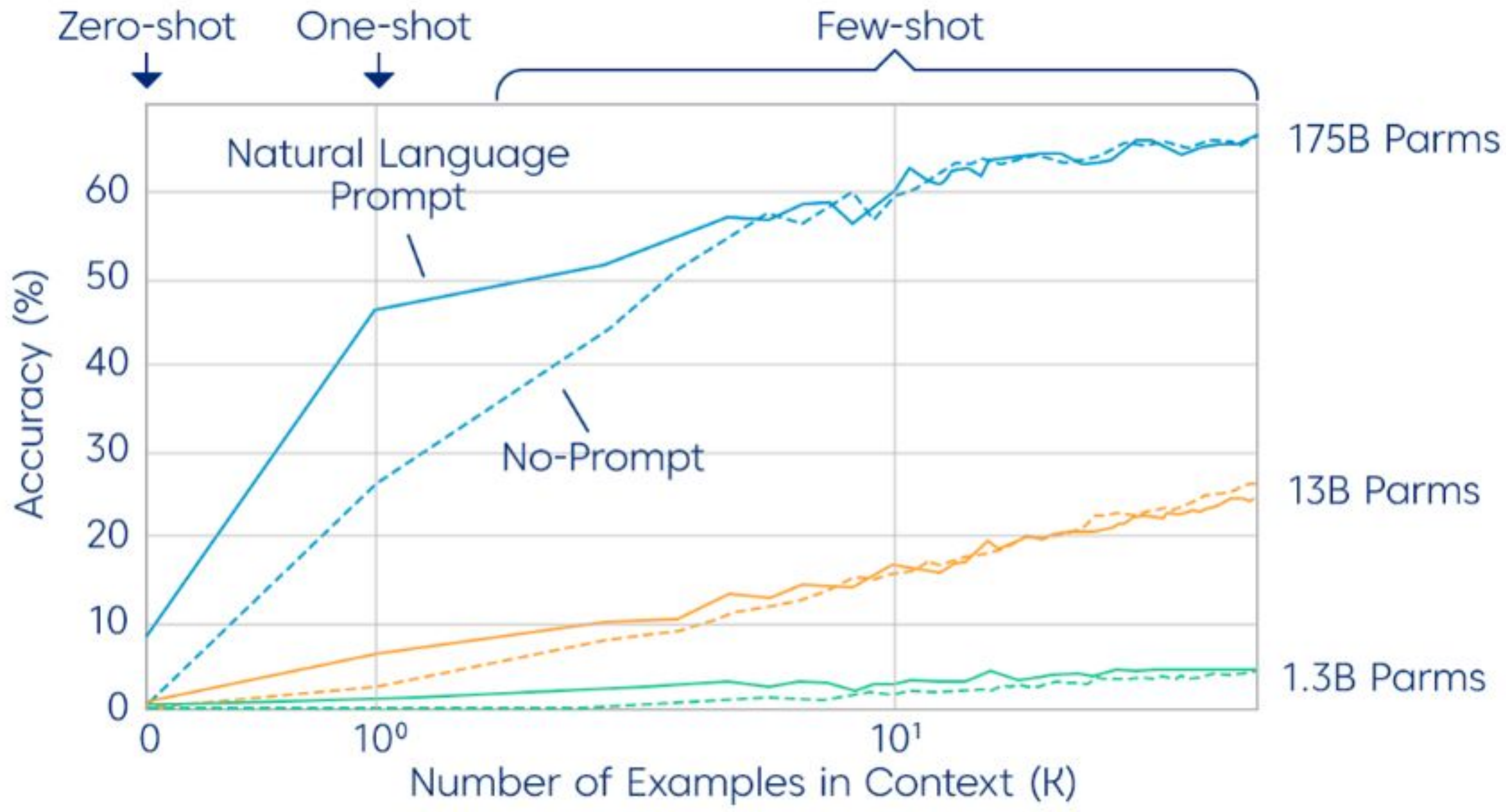
Artificial Intelligence : Large Language Models- GPT-3

Accionlabs



Dataset	Quantity (tokens)
Common Crawl (filtered)	410 billion
WebText2	19 billion
Books1	12 billion
Books2	55 billion
Wikipedia	3 billion

Larger models are learning efficiently from in-context information



SCIFORCE

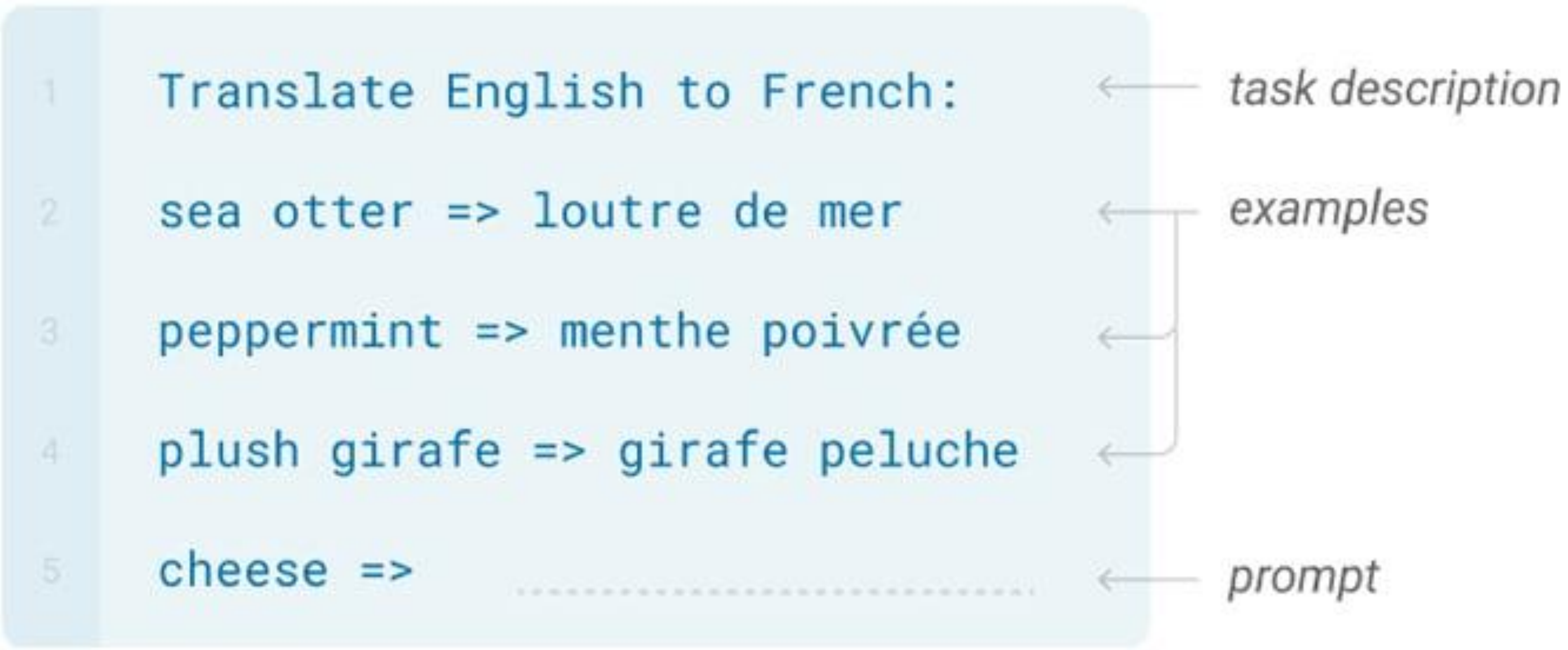
SCIFORCE

GPT-3 – Augmented learning : metadata and schema training

Accionlabs

Few-shot

In addition to the task description, the model sees a few examples of the task. No gradient updates are performed.



Query: What is the count of proteins?

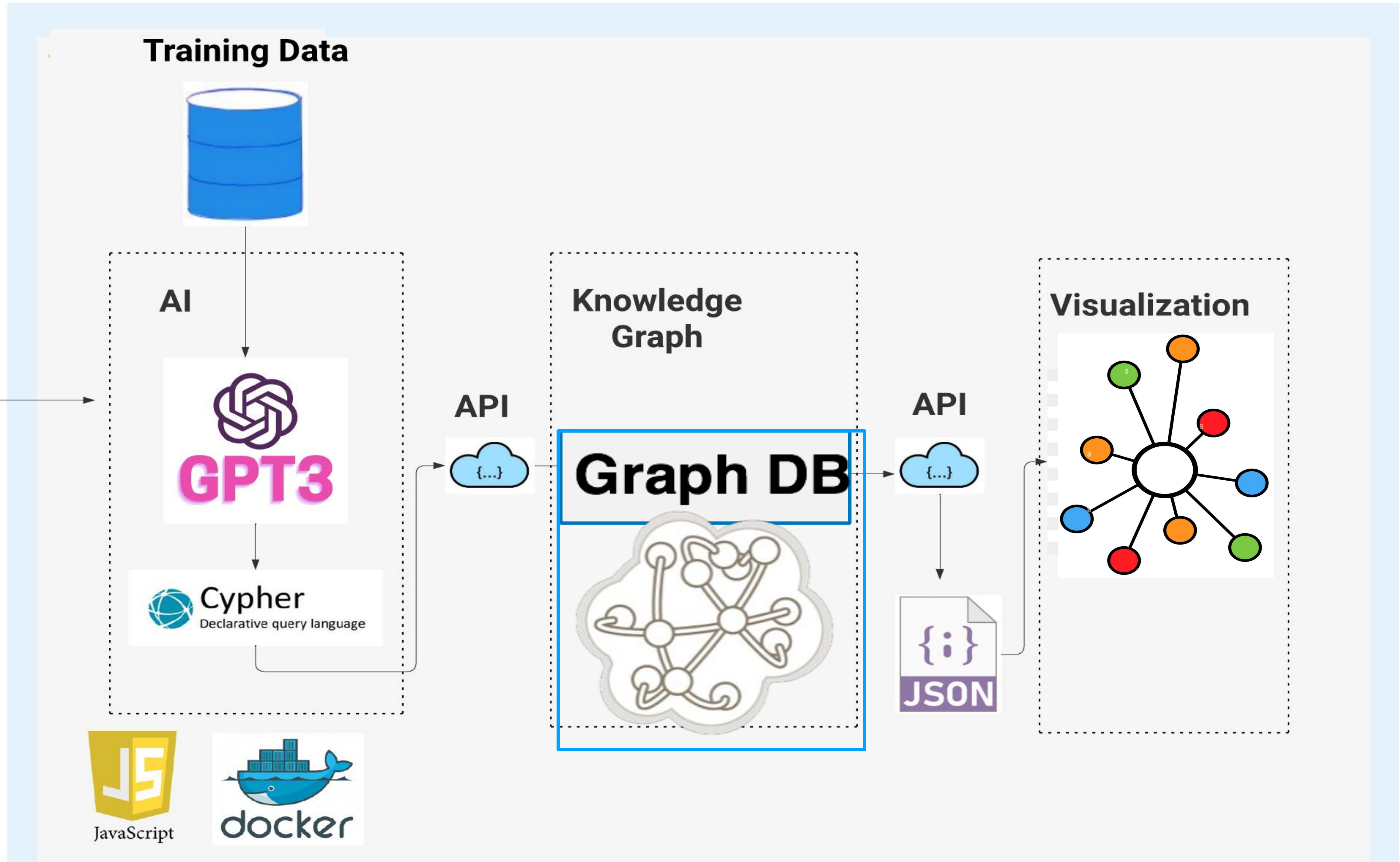
Output: MATCH (p:Protein) RETURN COUNT(DISTINCT p)

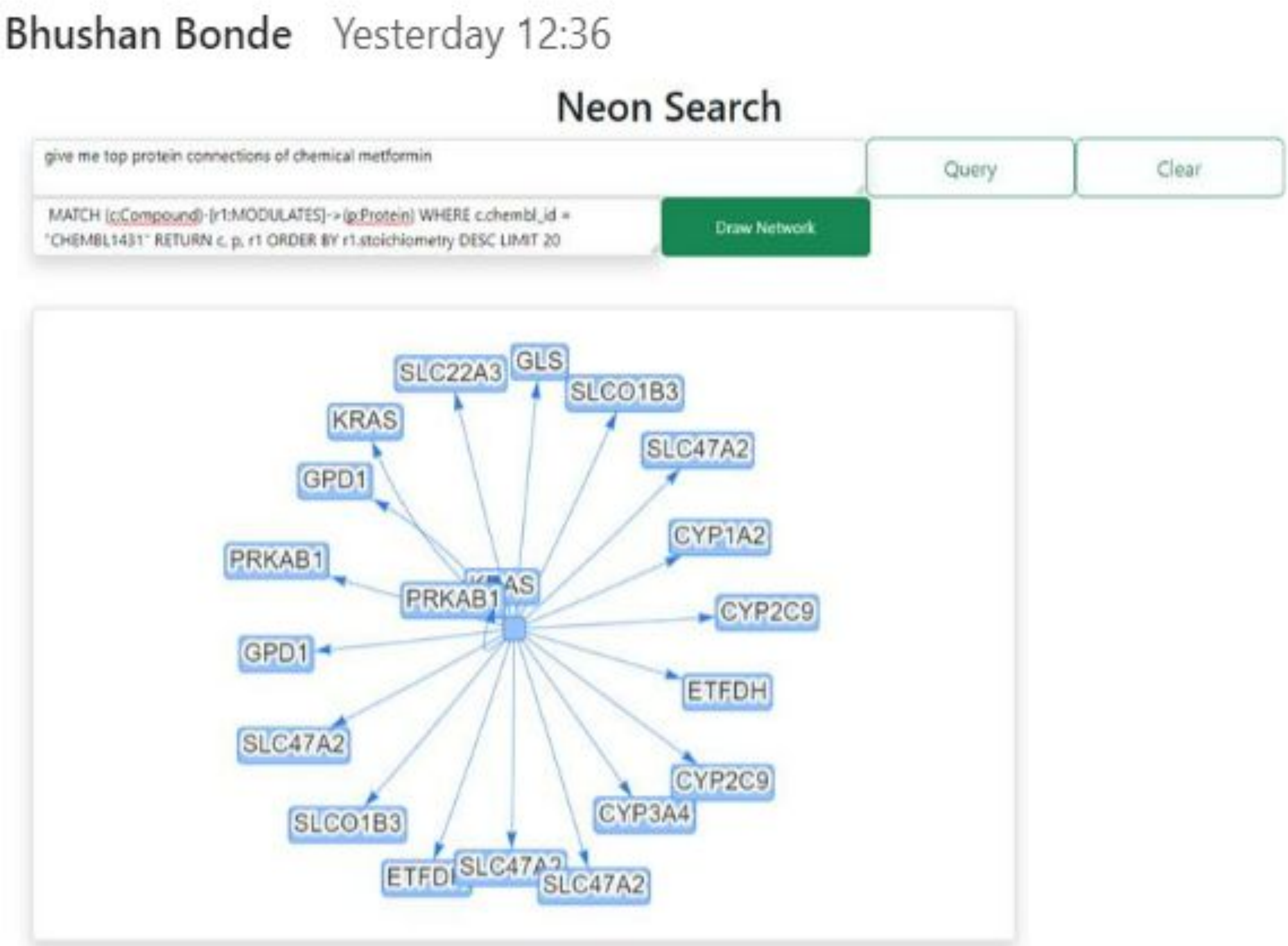
Query: What is the count of anatomy objects?

Output: MATCH (n:AnatomyObject) RETURN count(n)

GPT-3 – Augmented learning

Accionlabs





- **Privacy and IP preserving** layer was added by removing sensitive info ..eg, chemical names, IP sensitive info and GDPR compliance, using Data Anonymization tool

GPT-3 – Challenges and learning

Accionlabs

•ChatGPT sometimes writes **plausible-sounding** but **incorrect or nonsensical** answers.

Fixing this issue is challenging, as:

(1) during RL training, there's currently no source of (ground) truth;

(2) (overfitting) training cautiously causes it to decline questions that it can answer correctly;

(3) Interpolation vs extrapolation : supervised training misleads the model -> what the model knows, rather than what the it can interpret from the data

•ChatGPT can be a bit **sensitive** to tweaks to the input phrasing, given **one phrasing** of a question, the model can claim to not know the answer, but given a **slight rephrase**, can answer correctly.

•Biased toward **verbosity and overuses** certain phrases, biases in the training data

E.g. trainers prefer longer answers that look more comprehensive) and well-known over-optimization issues.

Accion INNOVATION SUMMIT 2023

Accionlabs

Summary

Knowledge Graphs are special graph

- Entities or Nodes: objects, events, labels or concepts ….
- Relationships: Interlinked by certain knowledge or relations between the entities…
- Knowledge Graphs (KG) enable efficient means of data management, storage, and retrieval
- Almost no data duplication, thereby providing the single snapshot of the data held by the organisation.
- The added benefit of Graphs are they allow efficient, impactful visualisation of the data.

Accionlabs

Accion INNOVATION SUMMIT 2023

Bhushan Bonde (bhushan.b@e-zest.com)

Special thanks to
Shrijeet Polke – E-Zest
Jaywant Deshpande – E-Zest
E-zest (Knowledge Graph/LLM) team