ROBOKOP and the Biomedical Data Translator: Efficiently Leveraging a Distributed Data Ecosystem

*Jim Balhoff[1], Chris Bizon[1], Steve Cox[1], **Karamarie Fecho**[1,2], Yaphet Kebede[1], Kenneth Morton[3], Alexander Tropsha[1,4], **Patrick Wang**.[3]*

[1]Renaissance Computing Institute (RENCI), University of North Carolina at Chapel Hill
[2]Copperline Professional Solutions
[3]CoVar Applied Technologies
[4]School of Pharmacy, University of North Carolina at Chapel Hill

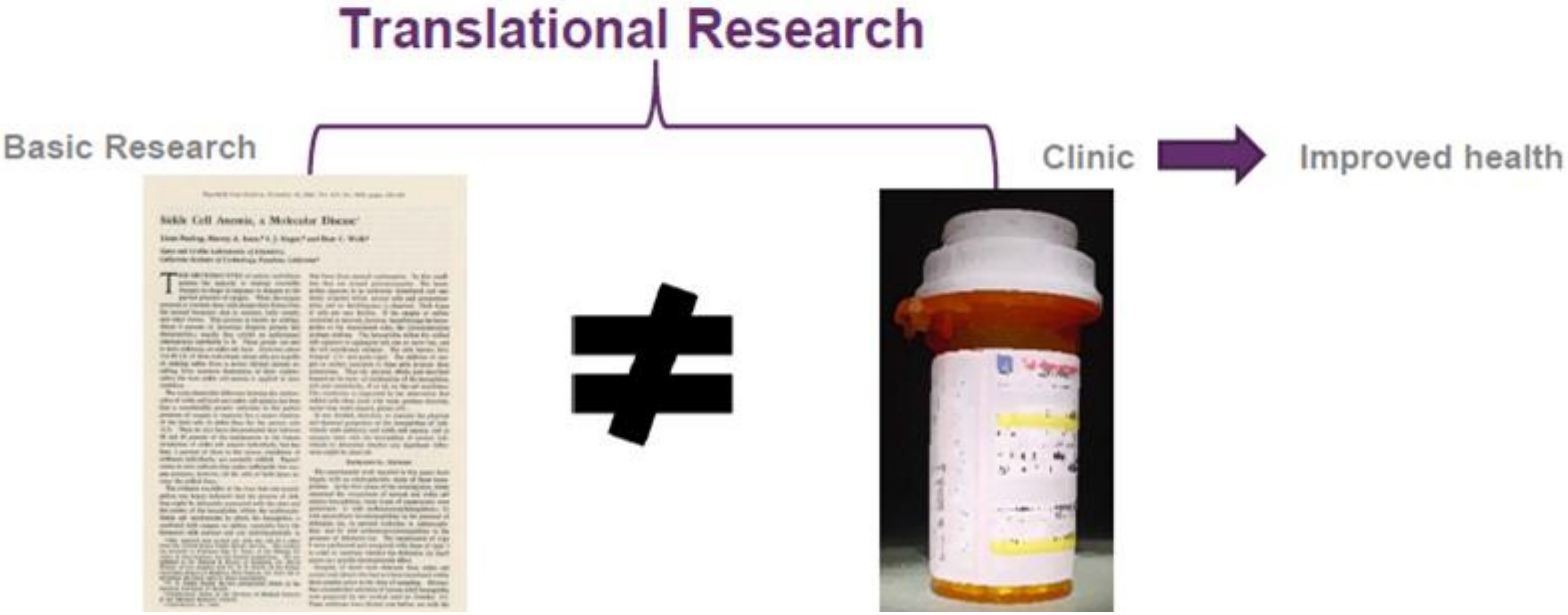
THE UNIVERSITY
of NORTH CAROLINA
at CHAPEL HILL

renci

NC TraCS
Institute

NIH National Center
for Advancing
Translational Sciences

NIH Mission…*is to seek fundamental knowledge about the nature and behavior of living systems and the application of that knowledge to enhance health, lengthen life, and reduce illness and disability*



Translational Research

Basic Research

≠

Clinic → Improved health

Image created by NCATS

NCATS Mission...*is to catalyze the generation of innovative methods and technologies that will enhance the development, testing and implementation of diagnostics and therapeutics across a wide range of human diseases and conditions*

Image created by NCATS

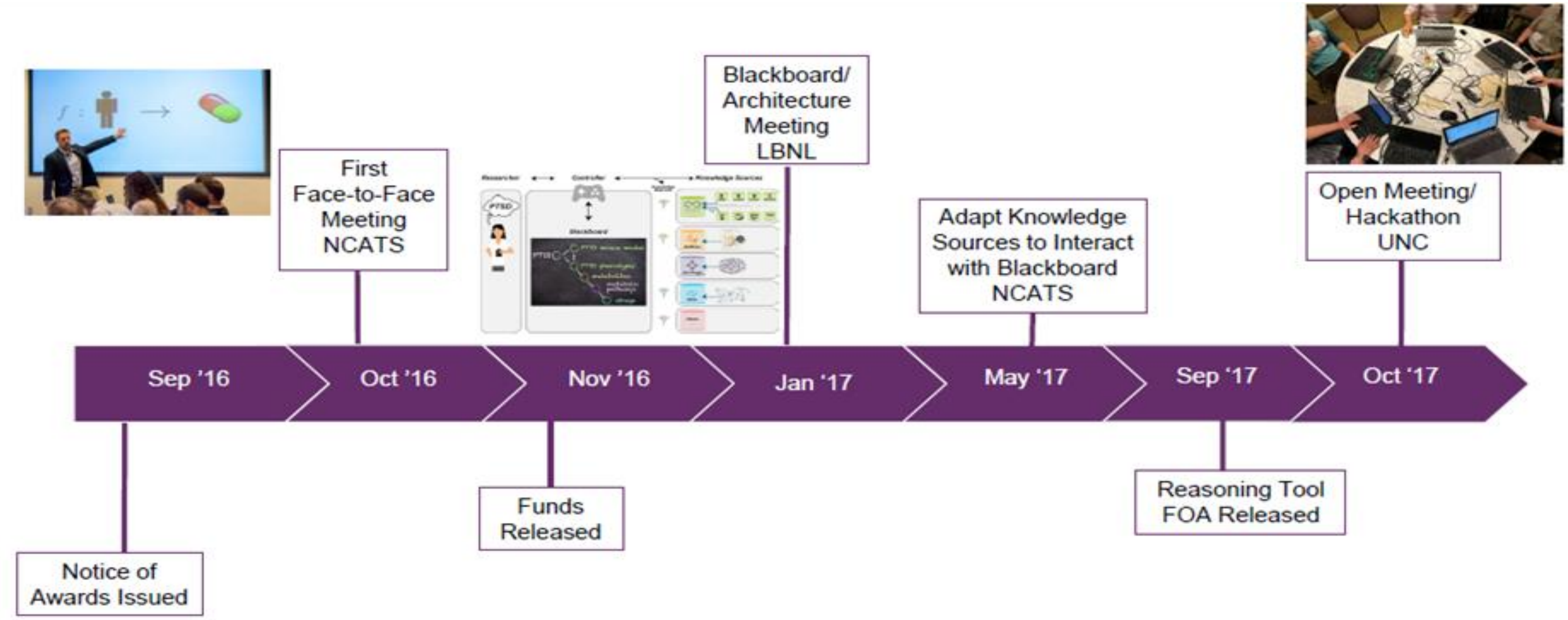
NCATS Biomedical Data Translator Program

- Overarching goal is to promote “**serendipitous**” discovery
- Primary goal is to build infrastructure to support and facilitate **data-driven** translational research on a large scale
- Essential aim is to **link as many datasets as possible** with one another and allow them to be **cross-queried** and **reasoned over** by translational researchers…

  …at **speed**, with **minimal barriers**, **scalability**, **accuracy**
- Fundamental tenet of the program is **open data** and **open collaborative team science**

NCATS Biomedical Data Translator Program

'Other Transaction Award (OTA)'



Sep '16 – Notice of Awards Issued

Oct '16 – First Face-to-Face Meeting NCATS

Nov '16 – Funds Released

Jan '17 – Blackboard/ Architecture Meeting LBNL

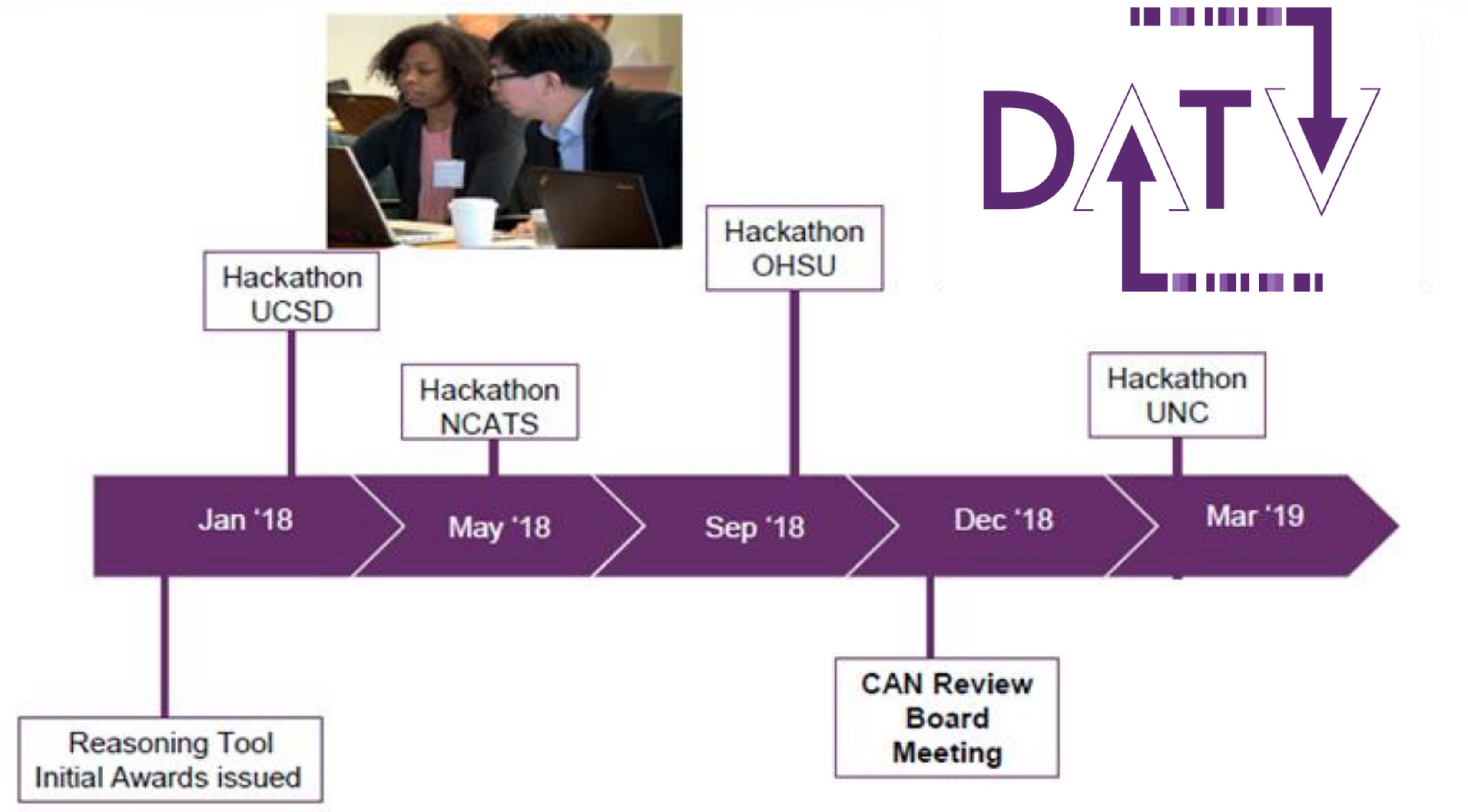
May '17 – Adapt Knowledge Sources to Interact with Blackboard NCATS

Sep '17 – Reasoning Tool FOA Released

Oct '17 – Open Meeting/ Hackathon UNC

Image created by NCATS

NCATS Biomedical Data Translator Program



Jan '18

Reasoning Tool Initial Awards issued

Hackathon UCSD

May '18

Hackathon NCATS

Sep '18

Hackathon OHSU

Dec '18

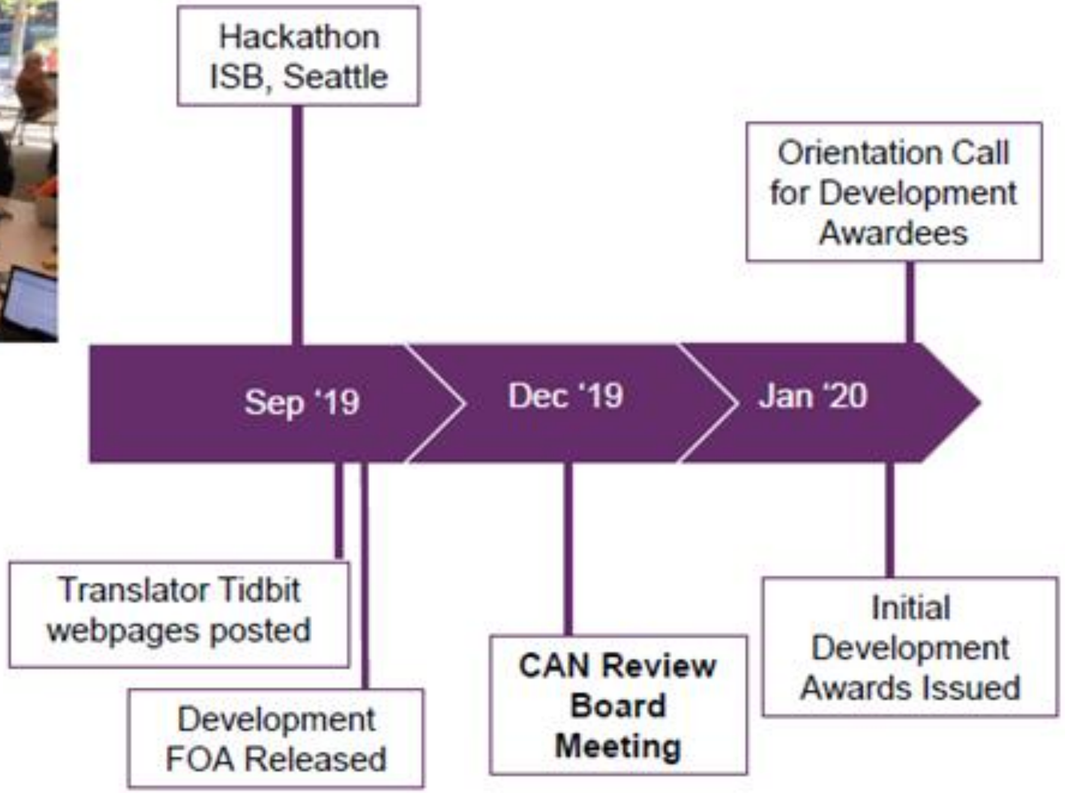
CAN Review Board Meeting

Mar '19

Hackathon UNC

Image created by NCATS and adapted here

NCATS Biomedical Data Translator Program



Hackathon ISB, Seattle

Orientation Call for Development Awardees

Sep '19

Dec '19

Jan '20

Translator Tidbit webpages posted

Development FOA Released

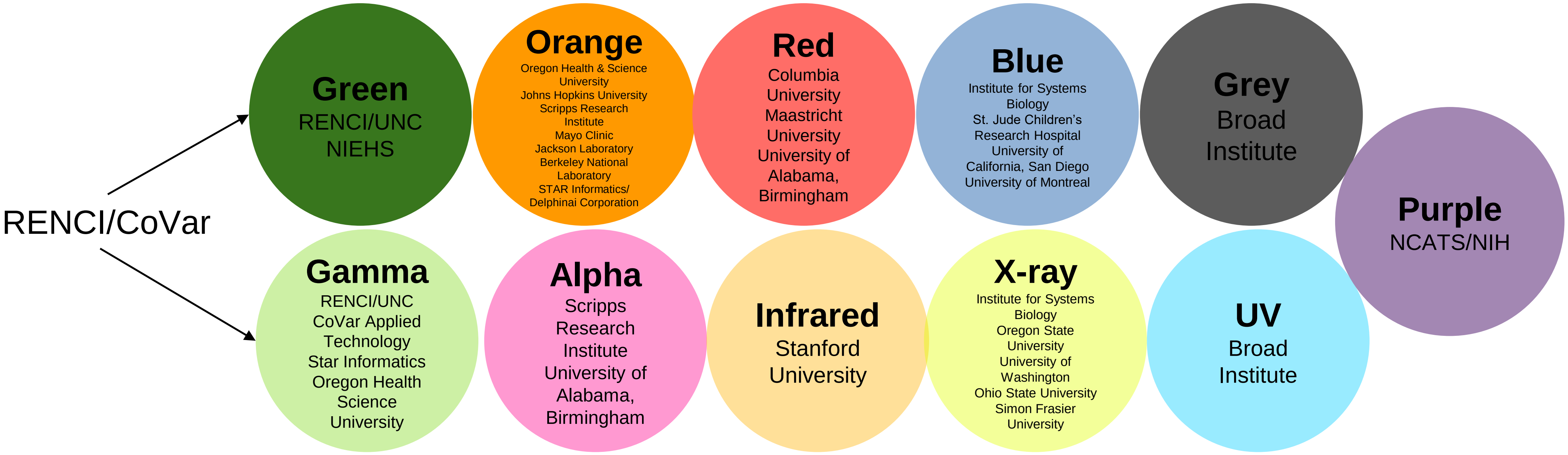
CAN Review Board Meeting

Initial Development Awards Issued

Image created by NCATS

The Biomedical Data Translator Consortium:
11 Teams, 28 institutions, ~200 team members

Phase I: 09/24/2016 - 12/31/2019



RENCI/CoVar

Green
RENCI/UNC
NIEHS

Orange
Oregon Health & Science University
Johns Hopkins University
Scripps Research Institute
Mayo Clinic
Jackson Laboratory
Berkeley National Laboratory
STAR Informatics/
Delphinai Corporation

Red
Columbia University
Maastricht University
University of Alabama, Birmingham

Blue
Institute for Systems Biology
St. Jude Children's Research Hospital
University of California, San Diego
University of Montreal

Grey
Broad Institute

Purple
NCATS/NIH

Gamma
RENCI/UNC
CoVar Applied Technology
Star Informatics
Oregon Health Science University

Alpha
Scripps Research Institute
University of Alabama, Birmingham

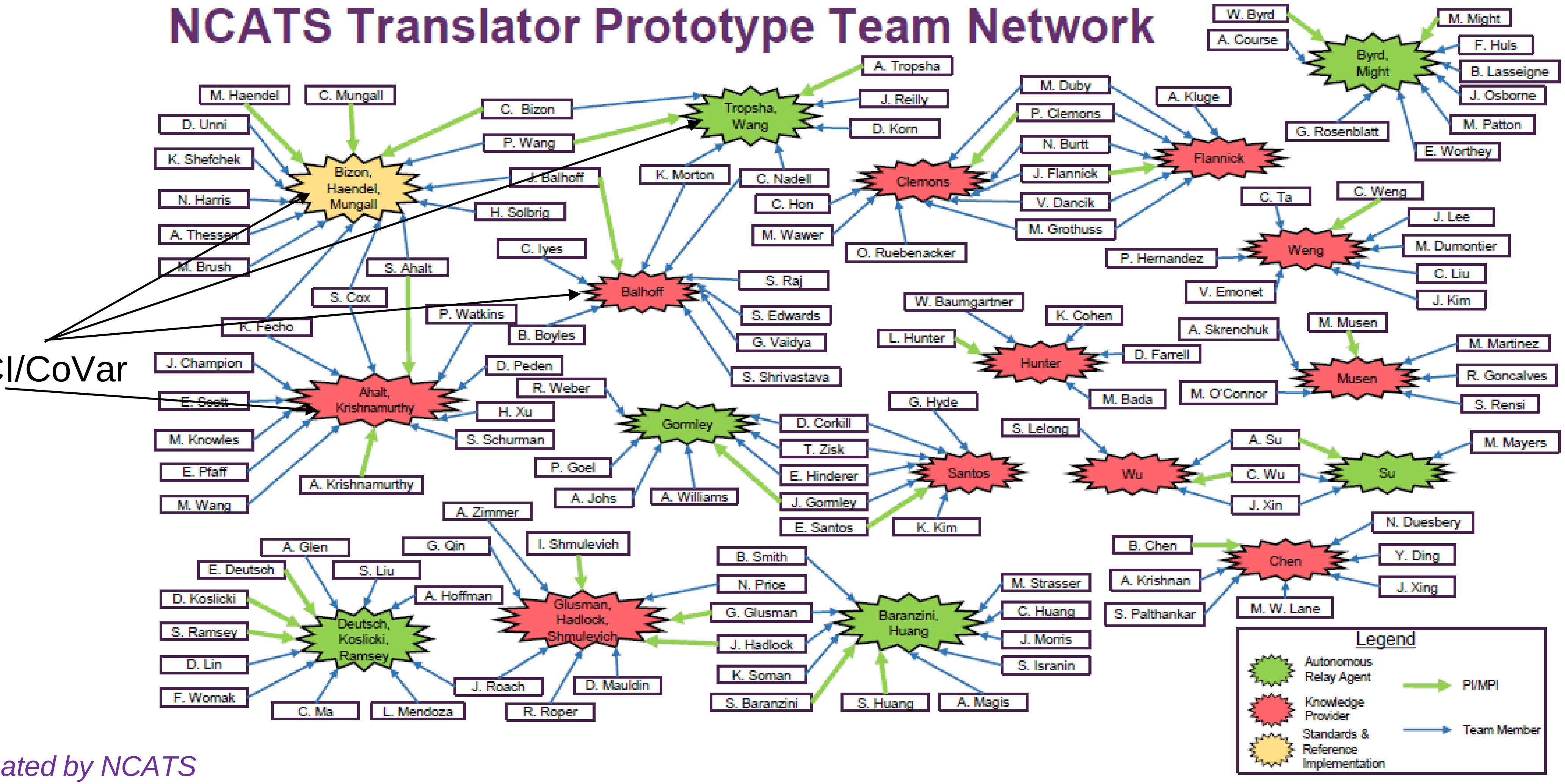
Infrared
Stanford University

X-ray
Institute for Systems Biology
Oregon State University
University of Washington
Ohio State University
Simon Frasier University

UV
Broad Institute

The Biomedical Data Translator Consortium:

Phase II: kick-off 01/27/2020



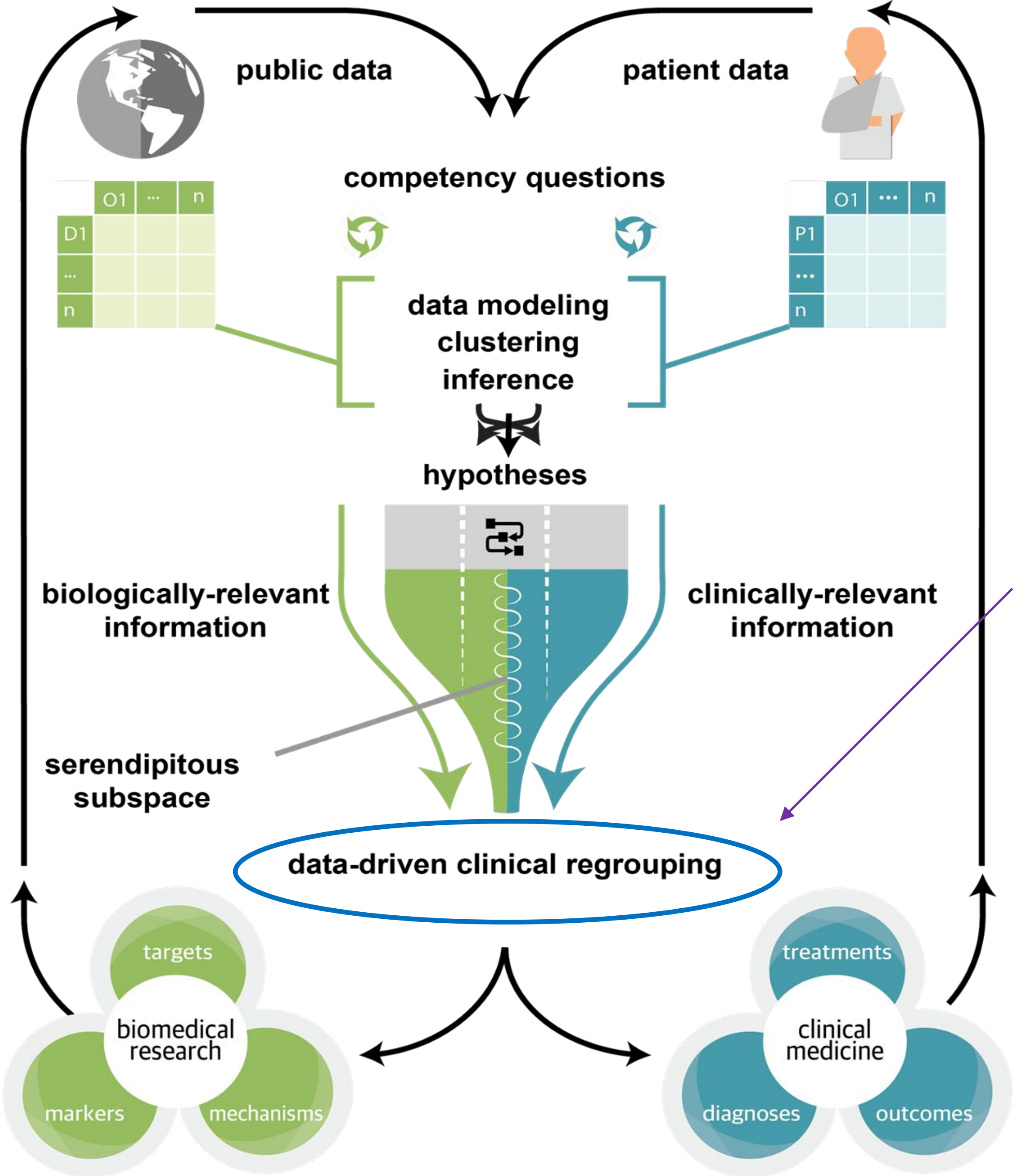
NCATS Translator Prototype Team Network

RENCI/CoVar

Image created by NCATS

Translator Vision

“Two hundred years ago, chemists created a comprehensive enumeration of the elements and systematic relationships among them. We envision the Translator doing the same for translational science.”
– Christopher P. Austin, MD, director of NCATS, with Christine M. Colvis, PhD, Noel T. Southall, PhD



Enhance (but not replace) human reasoning, catalyze translational research & drug discovery /repurposing

Image created by Julie McMurry, with input from the Biomedical Data Translator Consortium

Prototype Translator System Architecture

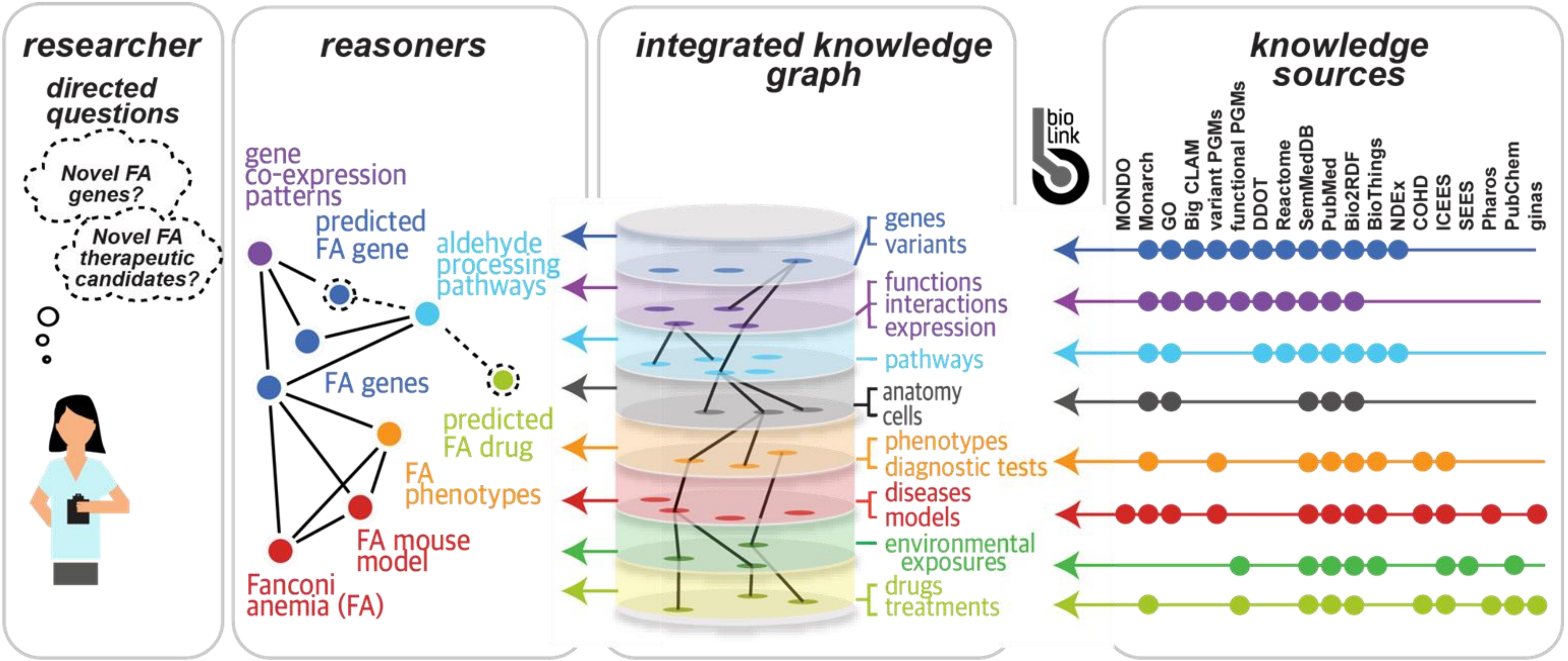


Image created by Julie McMurry, with input from the Biomedical Data Translator Consortium

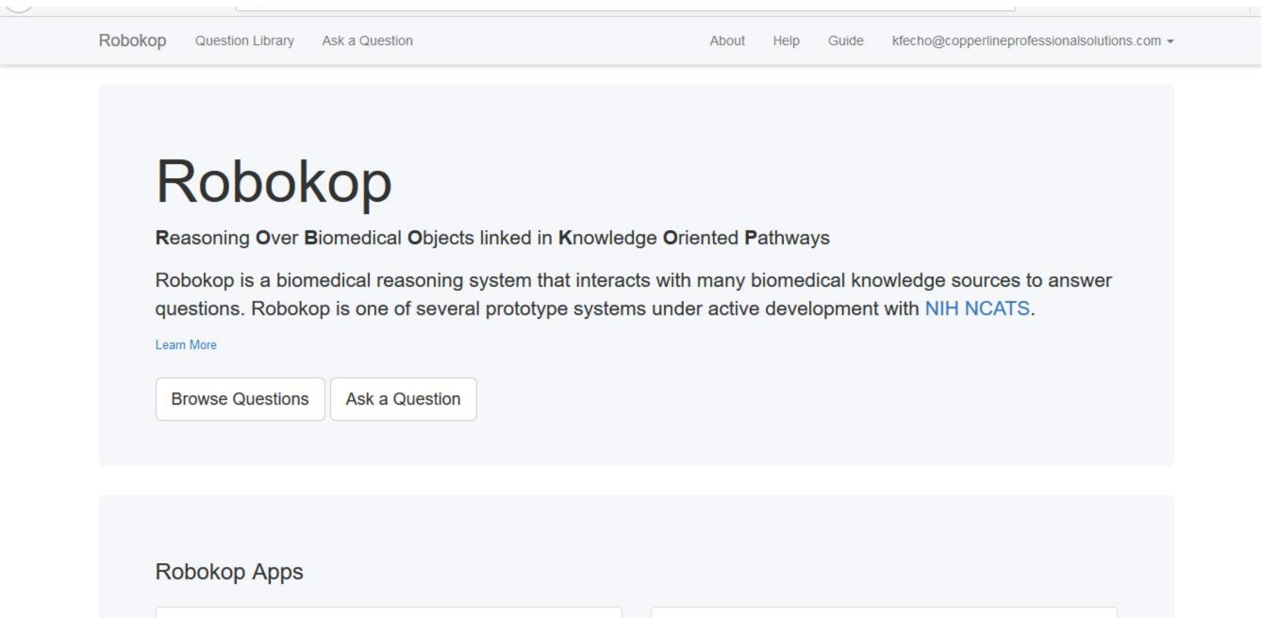
ROBOKOP (Reasoning Over Biomedical Objects linked in Knowledge-Oriented Pathways)

- An open knowledge graph (KG)–based question-answering system
- Can be used to derive mechanistic insights into real-world observations

what biological pathways might explain the association between exposure to carbon monoxide and multiple sclerosis?

how does hair dye influence susceptibility to breast cancer?

what genes might explain the effectiveness of isopropyl alcohol in the treatment of cyclic vomiting syndrome?



Robokop

Reasoning **O**ver **B**iomedical **O**bjects linked in **K**nowledge **O**riented **P**athways

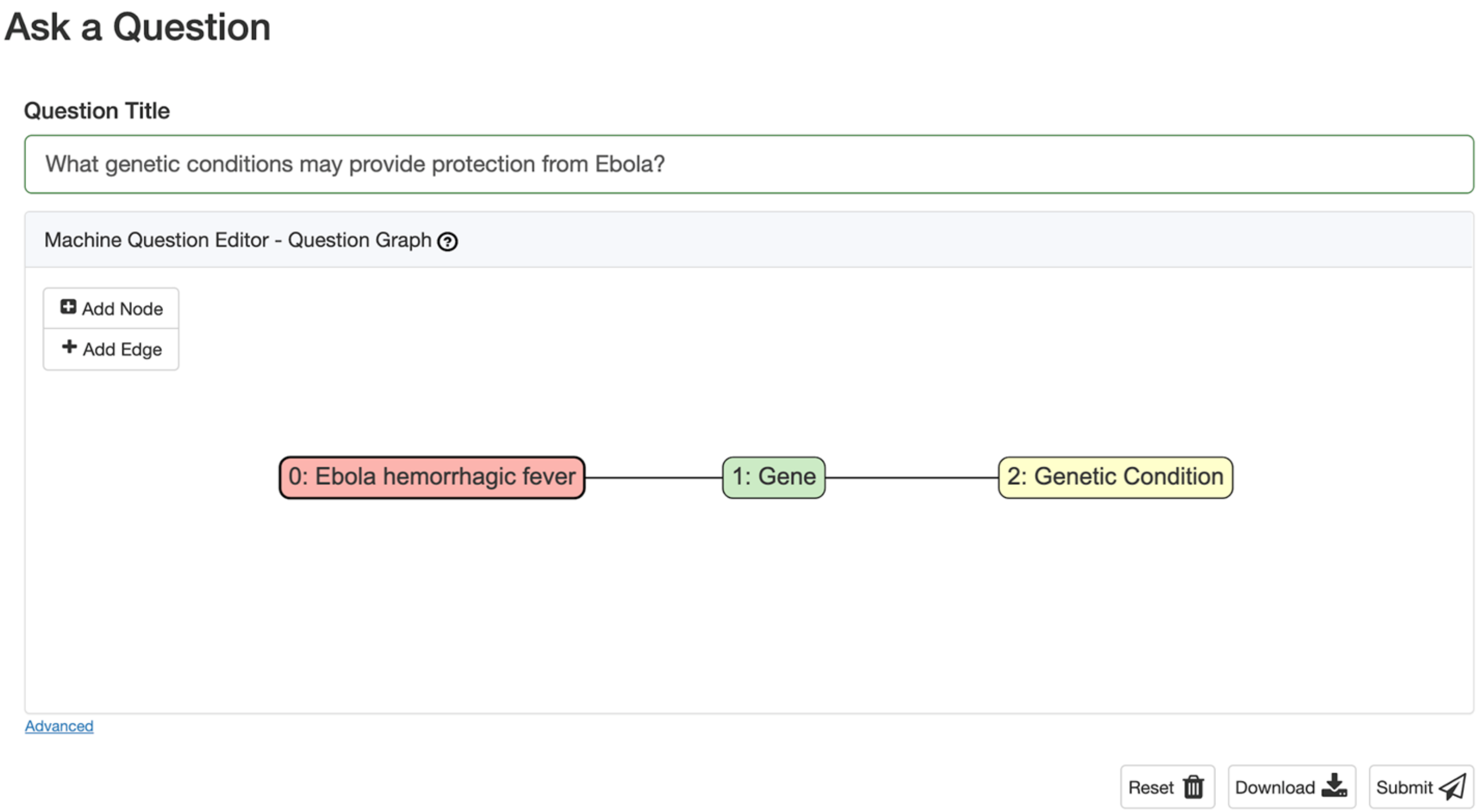
Robokop is a biomedical reasoning system that interacts with many biomedical knowledge sources to answer questions. Robokop is one of several prototype systems under active development with NIH NCATS.

Robokop Apps

robokop.renci.org

Demo

https://robokop.renci.org/



Ask a Question

Question Title

What genetic conditions may provide protection from Ebola?

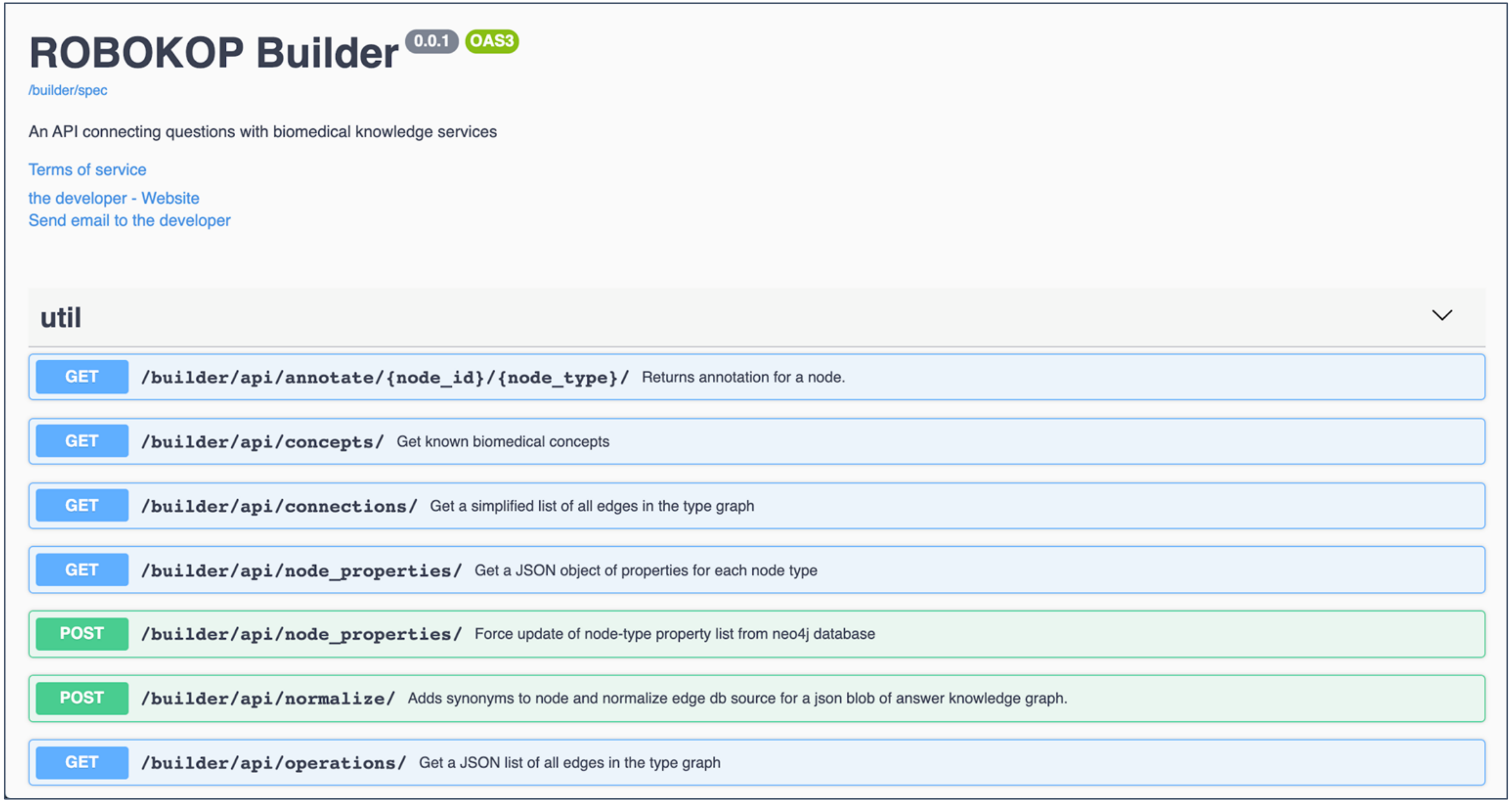
Machine Question Editor - Question Graph

Services

- “manager”
    - manages users
    - stores queries and results
    - presents custom web GUI
        - building and running queries
        - exploring results
- “ranker” / “messenger”
    - receives queries from public API, in JSON format
    - transpiles into Cypher query
    - collates, scores, and returns results
- “builder” / “interfaces”
    - consults registry of (third-party) APIs
    - identifies sequences of queries that could combine to produce answers
    - iteratively queries to gather relevant data
    - stores data in Neo4j database “robokopdb”
- “robokopDB”
    - publicly available Neo4j browser

Builder/Interfaces

https://robokop.renci.org/builder/apidocs/



ROBOKOP Builder

/builder/spec

An API connecting questions with biomedical knowledge services

Terms of service

the developer - Website

Send email to the developer

util

GET `/builder/api/annotate/{node_id}/{node_type}/` Returns annotation for a node.

GET `/builder/api/concepts/` Get known biomedical concepts

GET `/builder/api/connections/` Get a simplified list of all edges in the type graph

GET `/builder/api/node_properties/` Get a JSON object of properties for each node type

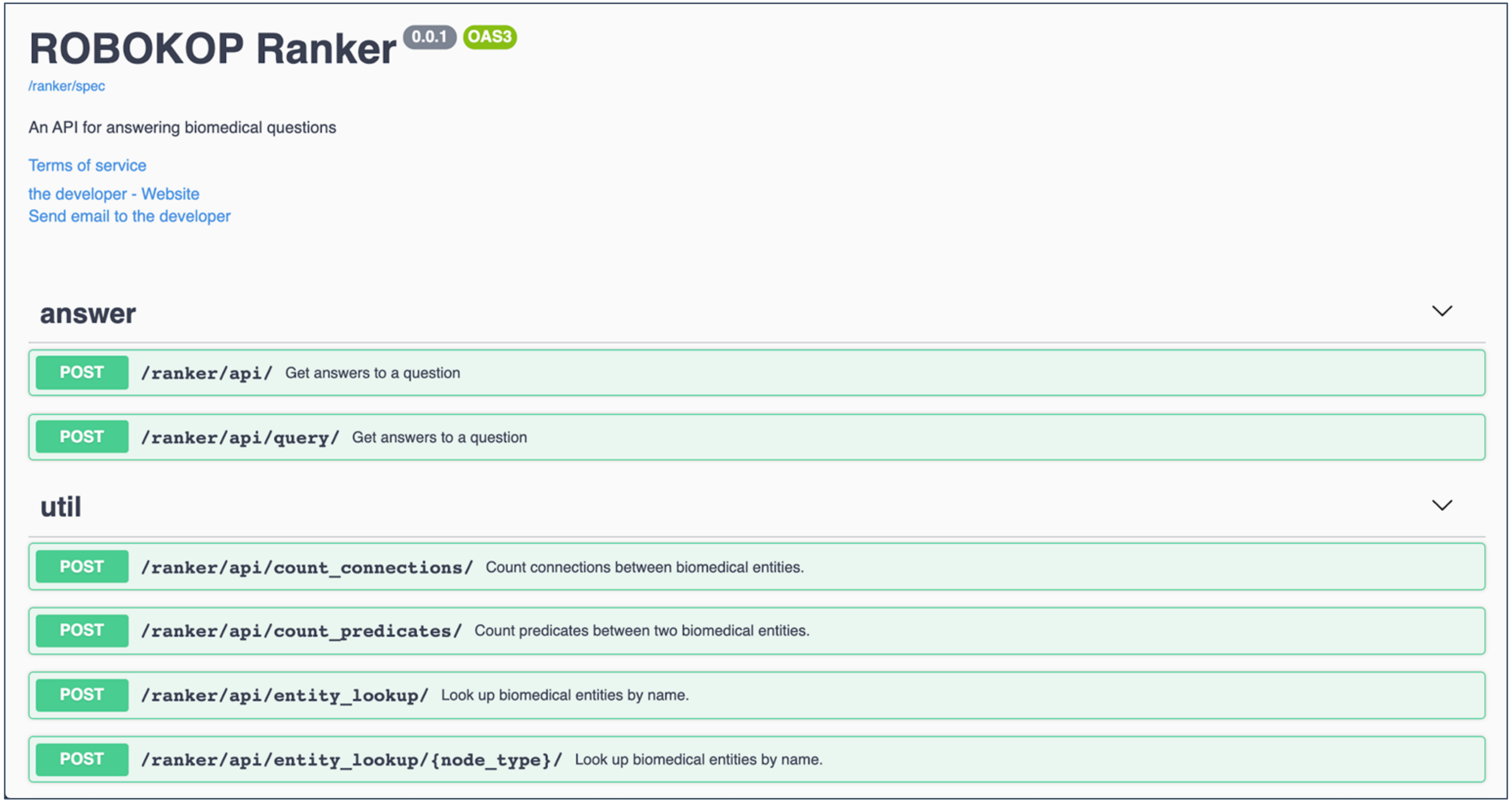
POST `/builder/api/node_properties/` Force update of node-type property list from neo4j database

POST `/builder/api/normalize/` Adds synonyms to node and normalize edge db source for a json blob of answer knowledge graph.

GET `/builder/api/operations/` Get a JSON list of all edges in the type graph

Ranker

https://robokop.renci.org/ranker/apidocs/



ROBOKOP Ranker

/ranker/spec

An API for answering biomedical questions

Terms of service

the developer - Website
Send email to the developer

answer

POST `/ranker/api/` Get answers to a question

POST `/ranker/api/query/` Get answers to a question

util

POST `/ranker/api/count_connections/` Count connections between biomedical entities.

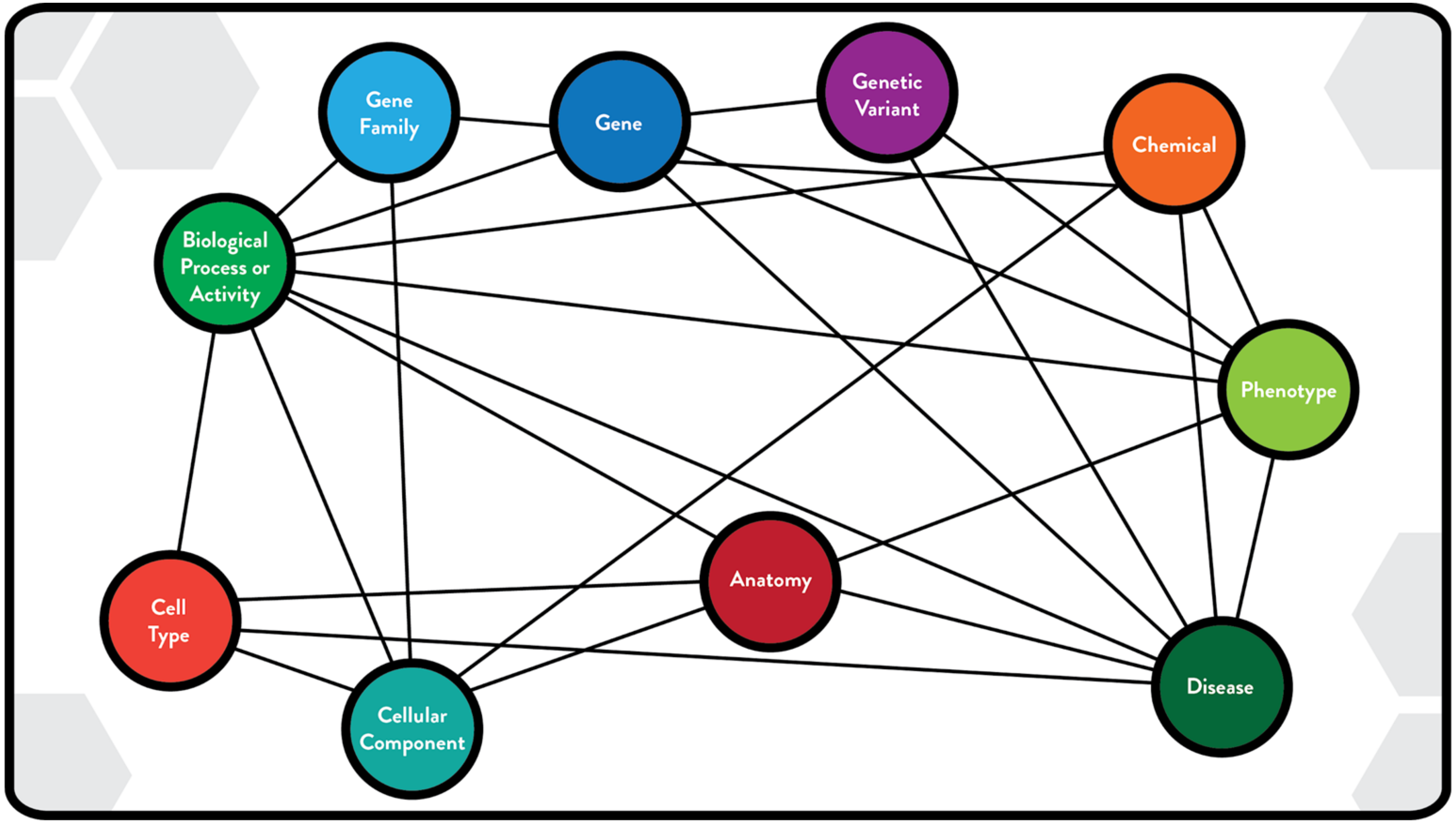
POST `/ranker/api/count_predicates/` Count predicates between two biomedical entities.

POST `/ranker/api/entity_lookup/` Look up biomedical entities by name.

POST `/ranker/api/entity_lookup/{node_type}/` Look up biomedical entities by name.

ROBOKOP KG

- > 6M nodes and 140M edges
- ~30 underlying biomedical data sources and bio-ontologies (e.g., KEGG, Monarch services, DrugCentral, Pharos, Comparative Toxicogenomics Database, Monarch Disease Ontology, Gene Ontology, ChEBI, Human Phenotype Ontology)
- Extensible to non-biomedical domains and knowledge sources



robokopkg.renci.org

ROBOKOP Tooling

- Supervisor: process control system
  - e.g. Flask process, Celery process, …
- Celery: distributed task queue
  - master Celery (Python) process
  - multiple Python worker processes
  - jobs distributed to workers via message broker (RabbitMQ)
  - results stored by workers in results database (Redis)

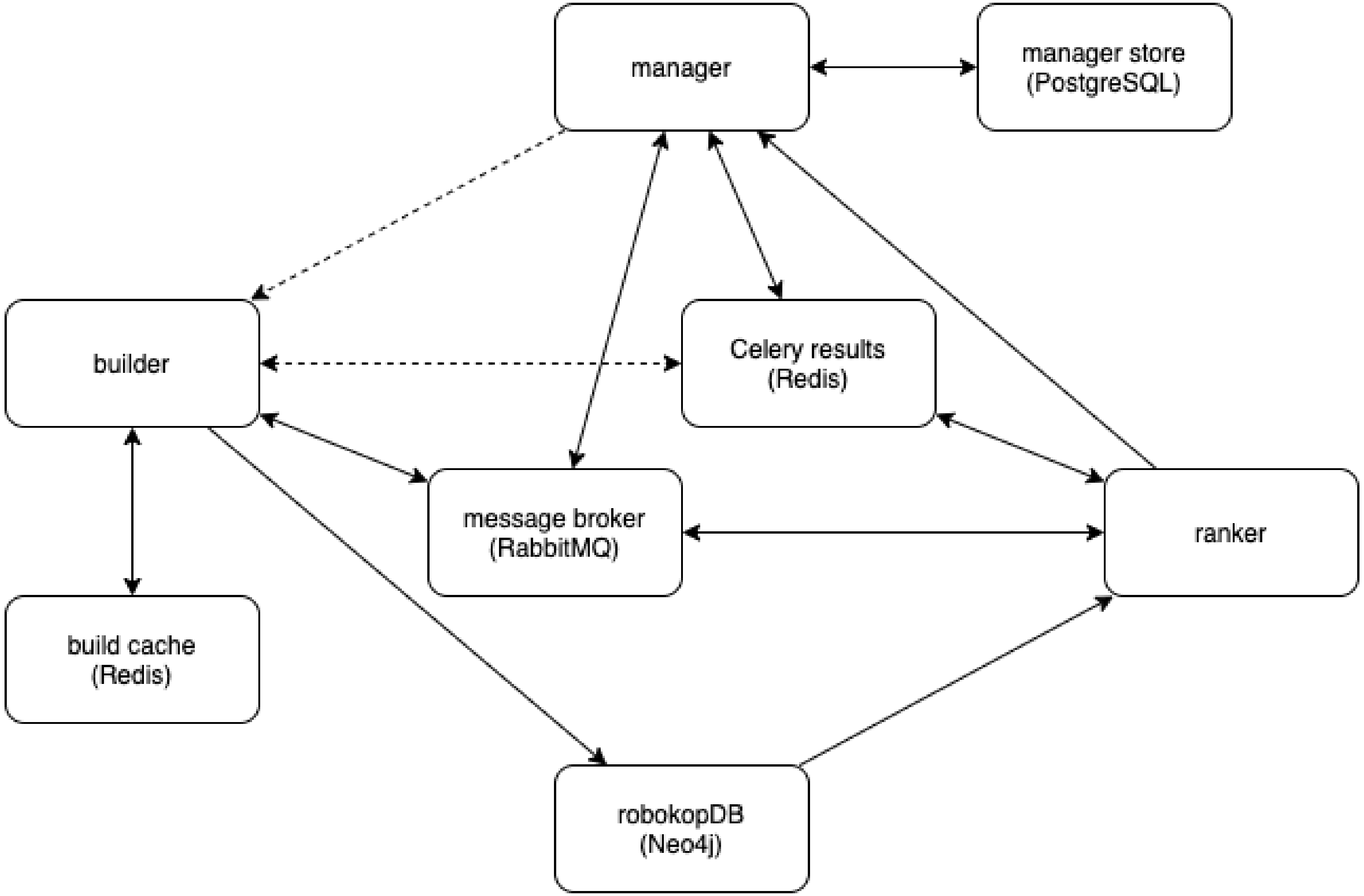
- NGINX, Gunicorn, Flask: web server
  - NGINX: reverse proxy
  - Gunicorn: WSGI HTTP server
    - multiple worker processes
  - Flask: web server framework
- Redis, Neo4j, PostgreSQL: caching
  - Third-party API calls (Redis)
  - Biomedical data (Neo4j)
  - Entire questions/answers (PostgreSQL)
- SQLAlchemy, Postgraphile + GraphQL
  - interface with PostgreSQL

Containers

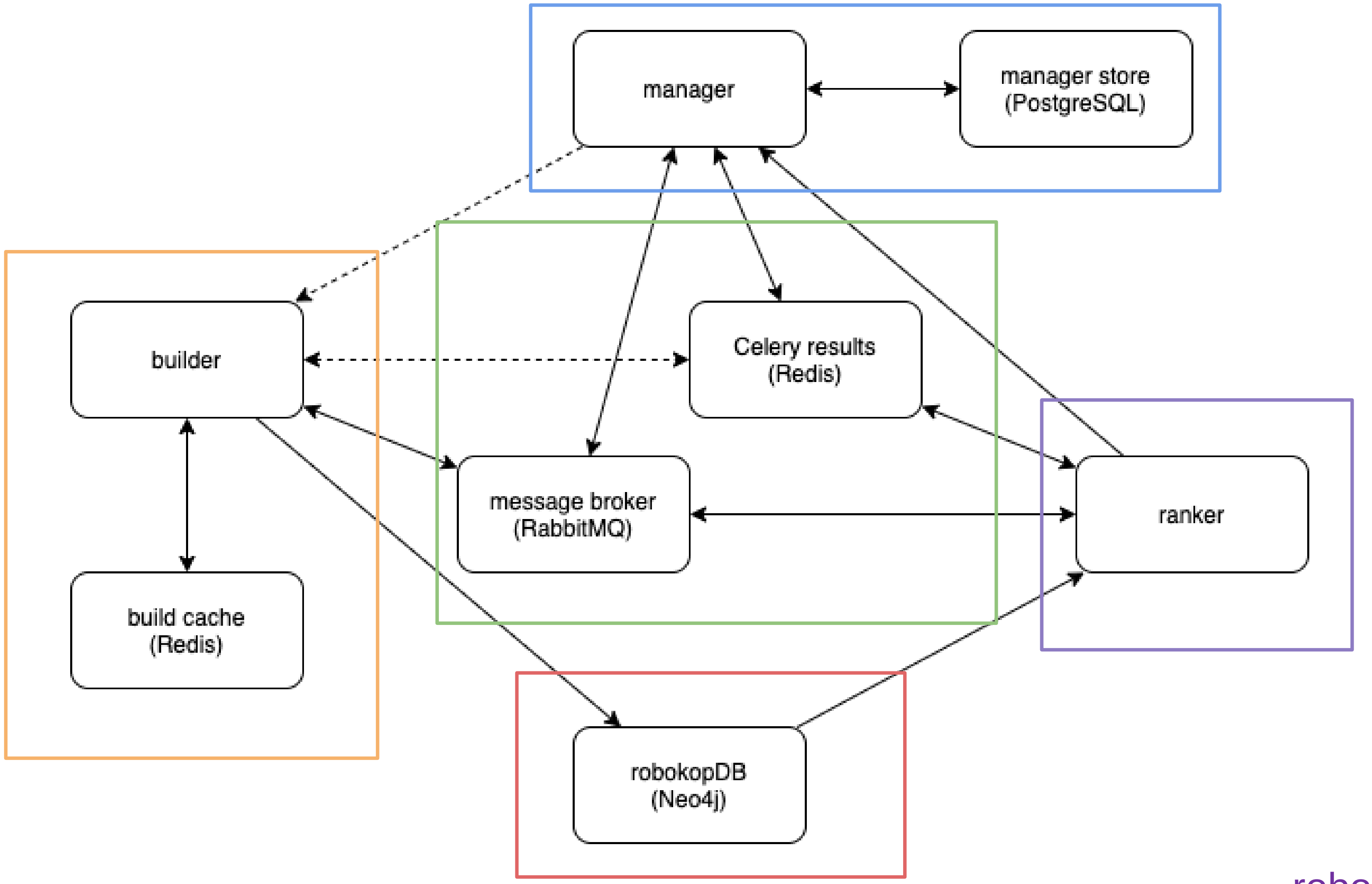
- manager
  - Python/Flask
- ranker/messenger
  - Python/Flask
- builder/interfaces
  - Python/Flask
- robokopDB
  - Neo4j

- reverse proxy
  - NGINX
- build cache
  - Redis
- manager store
  - PostgreSQL
- Celery results
  - Redis
- message broker
  - RabbitMQ

ROBOKOP Containers

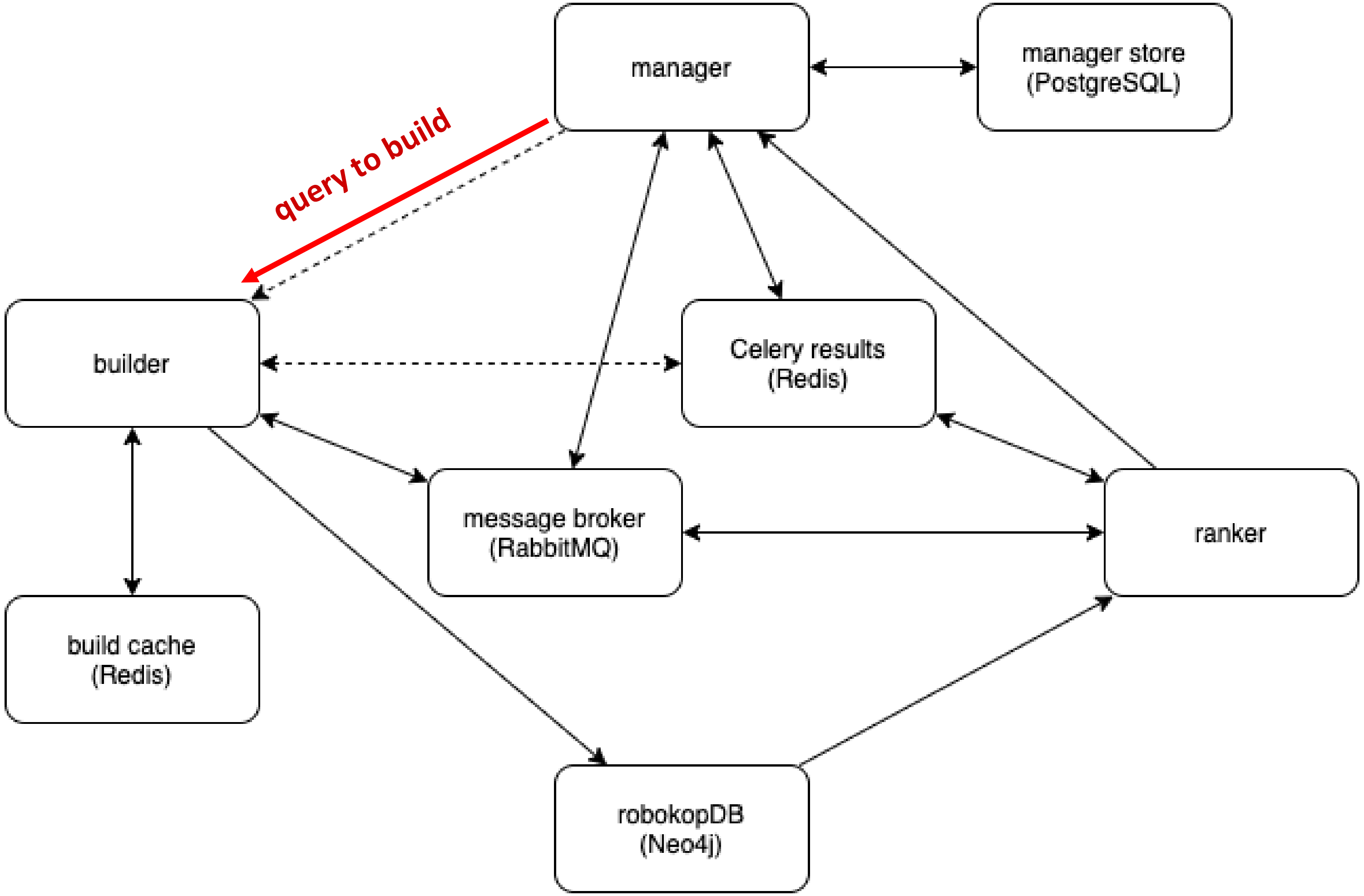


ROBOKOP Containers



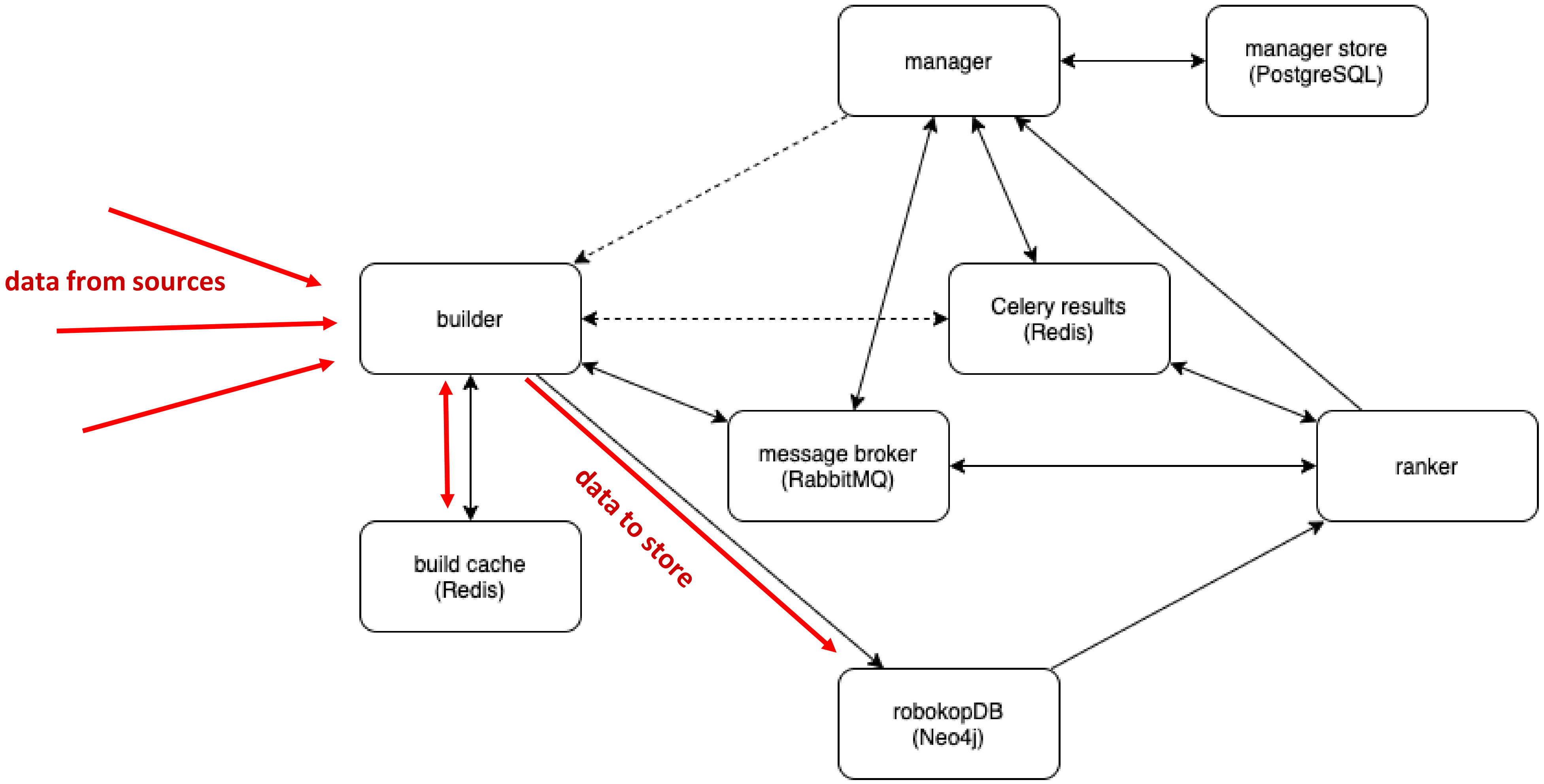
robokop.renci.org

ROBOKOP Containers



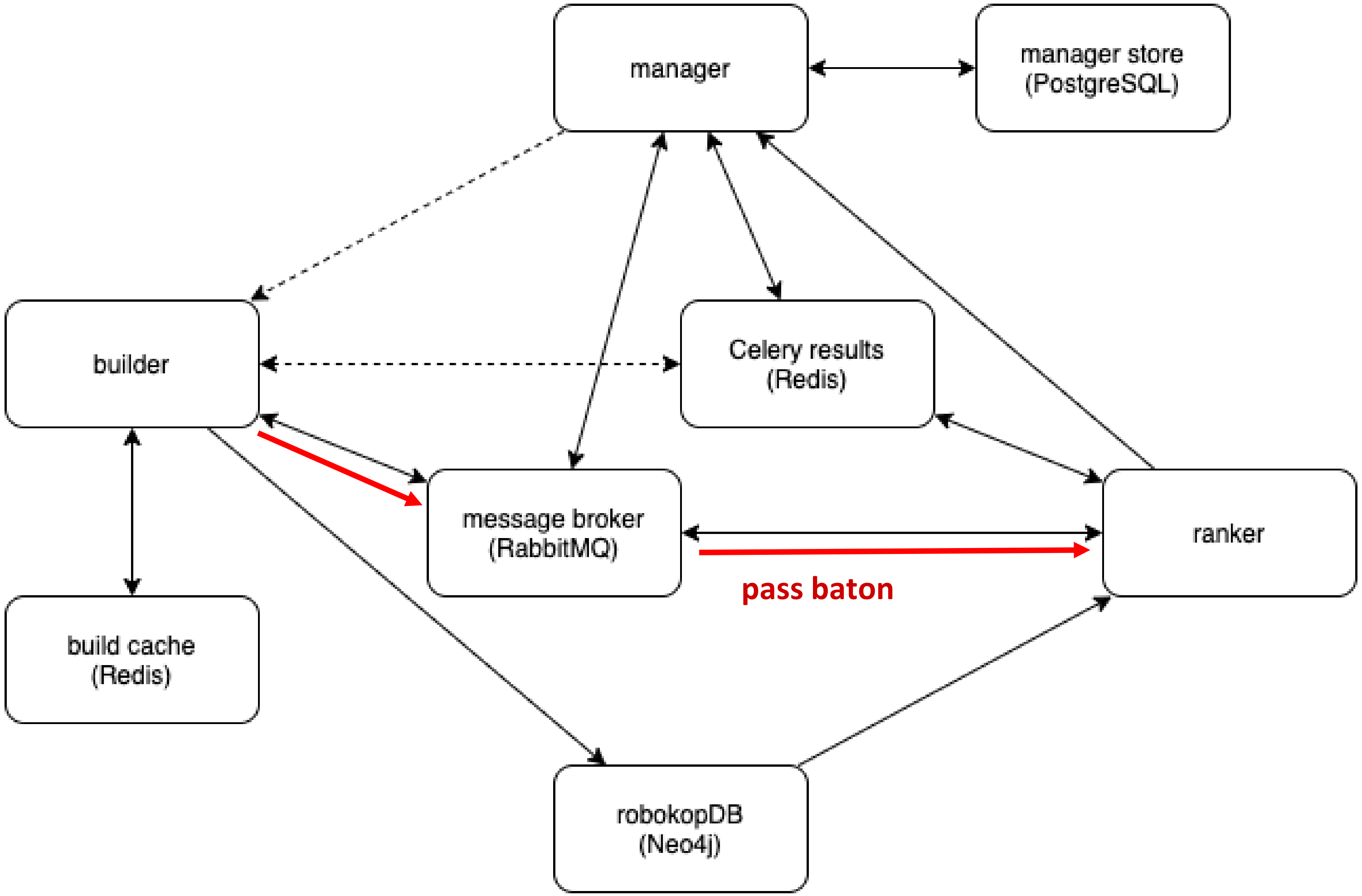
robokop.renci.org

ROBOKOP Containers



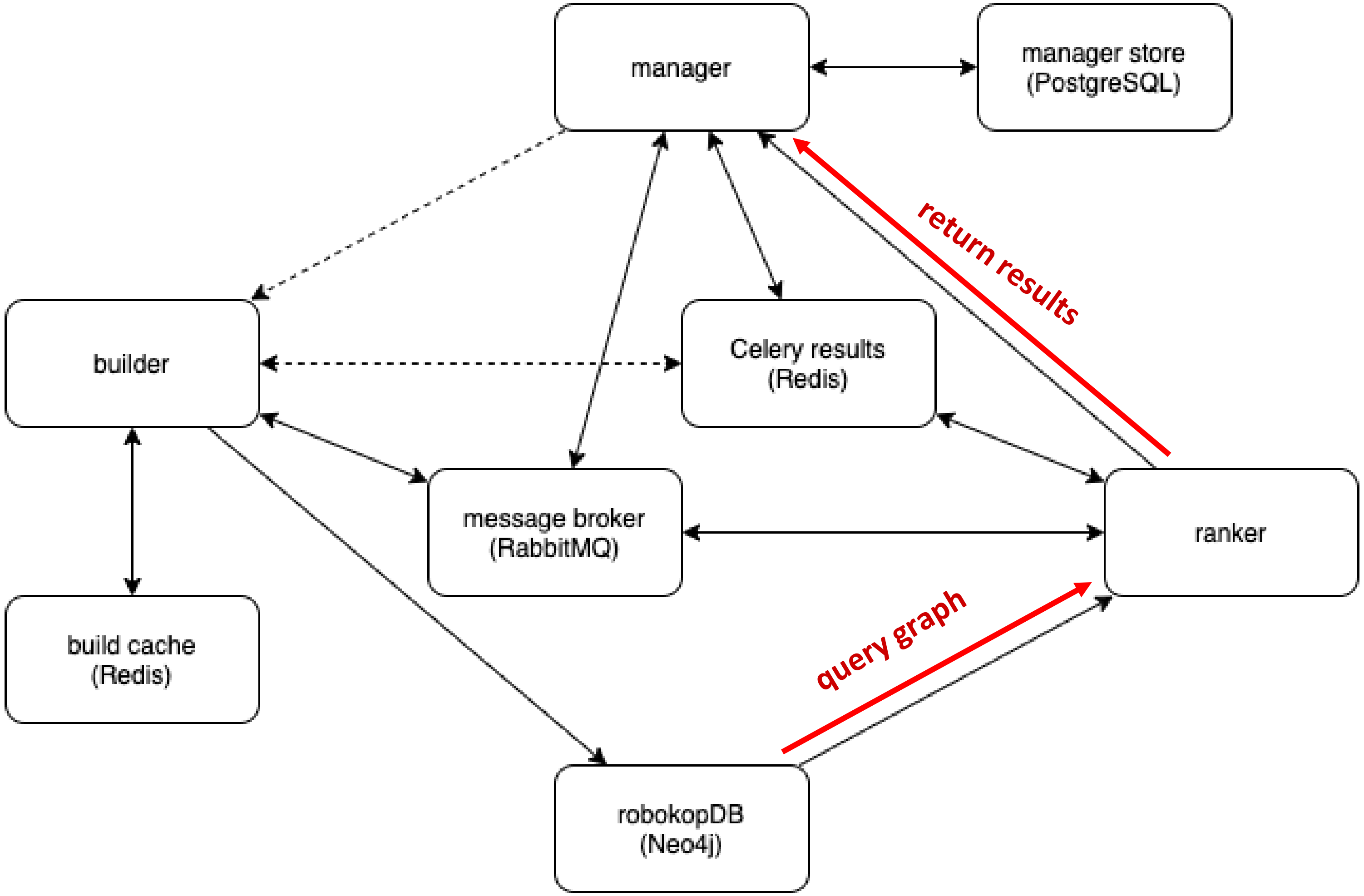
robokop.renci.org

ROBOKOP Containers

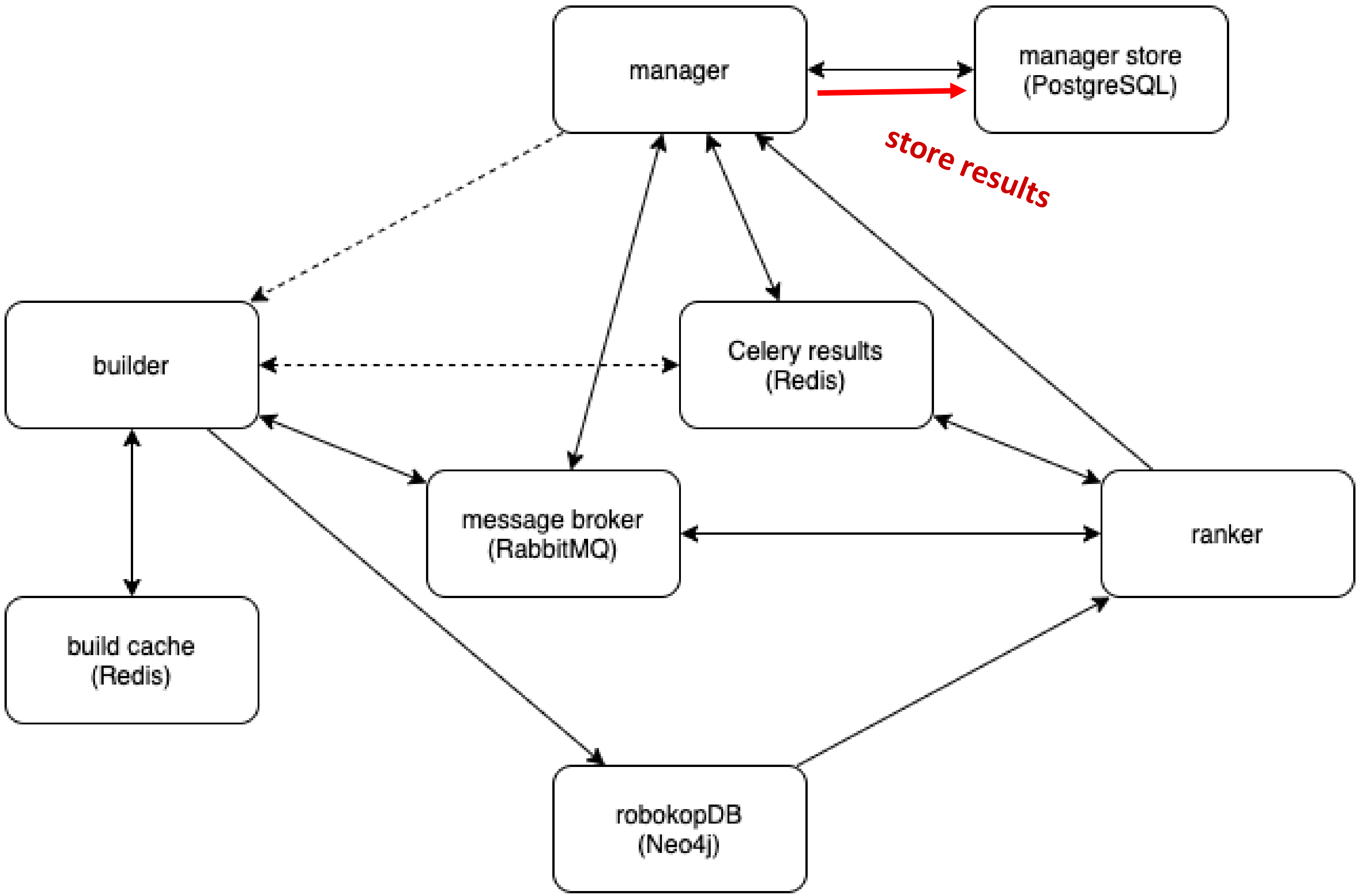


robokop.renci.org

ROBOKOP Containers



ROBOKOP Containers



robokop.renci.org

Shortcomings

- Brittle. If any of the following gets overworked, ROBOKOP becomes unusable:
  - robokopDB (Neo4j) - heap memory, disk space, ???
  - manager (Python) - SQLAlchemy, Celery workers
  - ranker (Python) - Celery workers
  - message broker (RabbitMQ) - *this has not happened yet!*
  - caches (Redis) - *this has not happened yet!*
  - manager store (PostgreSQL) - *this has not happened yet!*
- Opaque. Logging multiple processes, multiple containers is a mess
  - permissions for mapped Docker volumes on Linux are tricky

Future Work: ARAGORN

- Translator is moving from the 3-year “feasibility” phase into the 5-year “development” phase
- Greater focus on distributed system - want to avoid monolithic KG
- Want end-user interface to be responsive: seconds, not hours…
- No need to return all answers quickly, just the best ones…

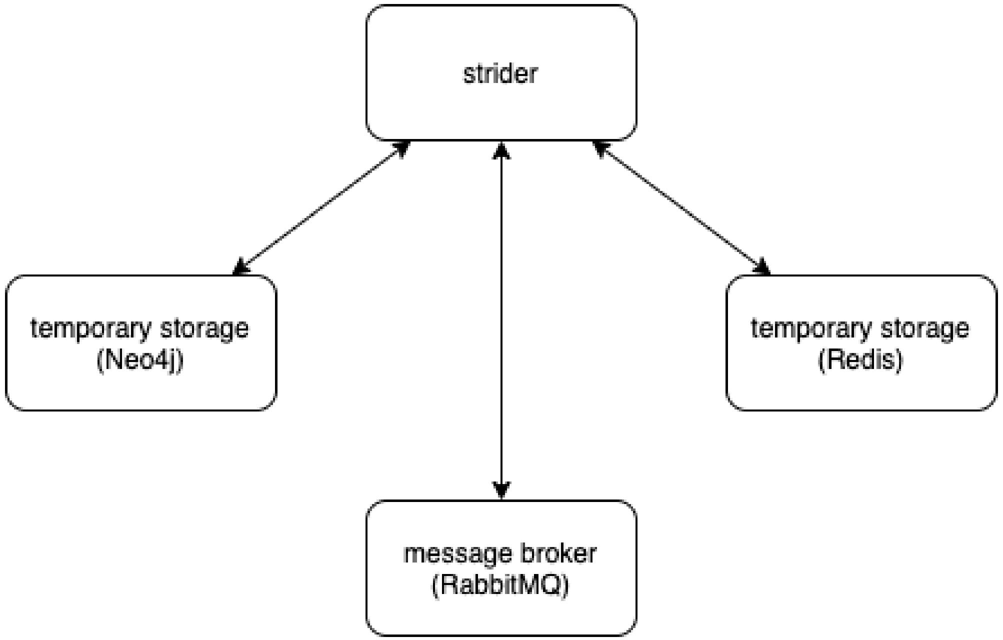
- Autonomous Relay Agent for Generation of Ranked Networks (ARAGORN)
  - a.k.a Strider
  - solution: asynchronous answering

ARAGORN

- “builder” and “ranker” get combined into a single component that builds and ranks iteratively to generate the best results first
- Plus novel algorithms for augmenting queries and “coalescing” results

- ROBOKOP was made efficient with caching and multiple processes
- ARAGORN will be efficient thanks to prioritization and asynchrony
  - We are largely I/O-bound, communicating with “third-party” Translator services

ARAGORN Containers



strider

temporary storage (Neo4j)

temporary storage (Redis)

message broker (RabbitMQ)

ARAGORN Tooling

- WSGI → ASGI
- Gunicorn → uvicorn
- Flask → Sanic/Starlette/FastAPI
  - asyncio, aioredis, aiormq, httpx, uvloop

- RabbitMQ + Redis for coordinating “threads”
- Neo4j for storing, querying biomedical data

Deployment

- ROBOKOP is essentially stateful
  - Relies on a single monolithic KG
  - Relies on long-term storage for questions/answers in PostgreSQL
  - Each component scales independently
- ARAGORN is at least *less* stateful
  - Relies on short-term storage using job-specific caches
  - Easier to scale horizontally - can replicate entire ARAGORN service and parallelize behind load balancer
    - e.g. Kubernetes

Gamma Team

- Jim Balhoff
- Mohit Bansal[†]
- Chris Bizon (co-PI)
- Richard Bruskiewich[†]
- Stephen Capuzzi
- Steven Cox
- Miles Crosskey[†]
- Karamarie Fecho
- Sudeep Mandal
- Kenneth Morton
- Eugene Muratov
- Kent Shefchek
- Andrew Thieme
- Alexander Tropsha (PI)
- Patrick Wang

Funding

Support was provided by the National Center for Advancing Translational Sciences, National Institutes of Health through the Biomedical Data Translator program (awards OT3TR002020 and OT2TR002514) and the Clinical and Translational Sciences Award program (award UL1TR002489).

Green Team

- Stanley Ahalt (PI)
- Sarav Arunachalam
- Stephen Appold
- Jim Balhoff
- Kira Bradford[†]
- James Champion
- Steve Cox
- Karamarie Fecho
- Karl Gustafson
- Adel Hanna[†]
- Ray Idaszak[†]
- Ann Moss Joyner[†]
- Ashok Krishnamurthy
- Benjamin Marsh[†]
- Asia Mieczkowska[†]
- Allan Parnell[†]
- Dave Peden
- Emily Pfaff
- Kimberly Robasky[†]
- Charles Schmitt
- Michael Stealey[†]
- Lisa Stillwell
- Alexander Tropsha (co-PI)
- Alejandro (Alex) Valencia
- Hao Xu
- Dongmei Yang[†]
- Hong Yi[†]

[†] Former team member

References and Resources

Green/Gamma Translator Documentation Website

- ROBOKOP page
- ROBOKOP APIs

Austin CP, Colvis CM, Southall NT. Deconstructing the translational tower of babel. *Clin Transl Sci* 2019;12(2):85. doi:10.1111/cts.12595.

Bizon C, Cox S, Balhoff J, Kebede Y, Wang P, Morton K, Fecho K, Tropsha A. ROBOKOP KG and KGB: Integrated Knowledge Graphs from Federated Sources. *J Chem Inf Model* 2019 Dec 23;59(12):4968–4973. doi: 10.1021/acs.jcim.9b00683.

Morton K, Wang P, Bizon C, Cox S, Balhoff J, Kebede Y, Fecho K, Tropsha A. ROBOKOP: An Abstraction Layer and User Interface for Knowledge Graphs to Support Question Answering. *Bioinformatics* 2019;pii:btz604. doi: 10.1093/bioinformatics/btz604. [Epub ahead of print]

The Biomedical Data Translator Consortium. The Biomedical Data Translator program: conception, culture, and community. *Clin Transl Sci* 2019;12(2):86–90. doi:10.1111/cts.12592.

The Biomedical Data Translator Consortium. Toward a universal biomedical data translator. *Clin Transl Sci* 2019;12(2):91–94. doi:10.1111/cts.12591.

Contact information:

- kfecho@copperlineprofessionalsolutions.com
- patrick@covar.com