49th International Conference on Very Large Data Bases

Effective Entity Augmentation By Querying External Data Sources

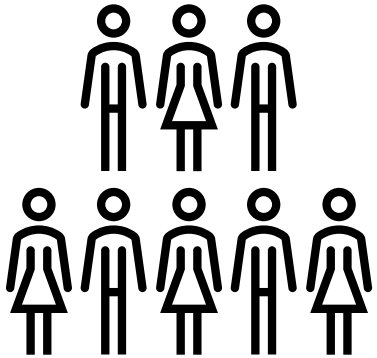
Christopher Buss, Jasmin Mousavi, Mikhail Tokarev, Arash Termehchy, David Maier, Stefan Lee

Oregon State University

Portland State UNIVERSITY

[Rare Daily Staff 2023] & [Ashburn & Thor 2004]

Drug Repositioning Can Save Lives



Patients with Castleman's disease

- Rare disease
- Potentially fatal: causes <u>severe inflammation</u>
- No effective treatments currently exist

Unfortunate reality:

Too rare: no financial incentive for companies to develop treatments

Alternative:

Find an existing drug to treat Castleman's disease

[Rare Daily Staff 2023] & [Ashburn & Thor 2004]

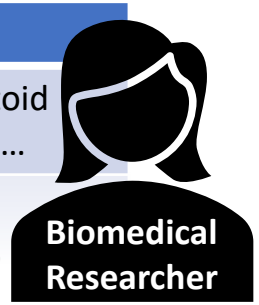
Identify a Candidate Drug

Find a candidate drug

FDA-Approved Drugs

brand_name	class	uses
Humira	TNF inhibitor	rheumatoid arthritis …
Enbrel	TNF inhibitor	plaque psoriasis

Local Data Source



Castleman's causes severe inflammation…

Humira is used to treat conditions involving severe inflammation

Candidate drug: Humira

Next step: gather more information about Humuria:

- Will it help or hurt?

[Rare Daily Staff 2023] & [Ashburn & Thor 2004]

Find External Sources

Local entity:

brand_name	class	uses
Humira	TNF inhibitor	rheumatoid arthritis …

FDA-Approved Drugs

brand_name	class	uses
Humira	TNF inhibitor	rheumatoid arthritis …
Enbrel	TNF inhibitor	plaque psoriasis

Local Data Source

Generic Drugs

generic_name	adverse_effects
Adalimumab	After treatment with adalimumab …
Etanercept	Etanercept binds specifically to tumor …

External Data Source

Bio Compounds

formula	mechanisms
$C_{6428}H_{9912}N_{1694}O_{1987}S_{46}$	Binds with specificity to tumor …
$C_{2224}H_{3475}N_{621}O_{698}S_{36}$	There are two distinct receptors …

External Data Source

[Rare Daily Staff 2023] & [Ashburn & Thor 2004]

What we Want: Info Relevant to Humira

Local entity:

brand_name	class	uses
Humira	TNF inhibitor	rheumatoid arthritis …

FDA-Approved Drugs

brand_name	class	uses
Humira	TNF inhibitor	rheumatoid arthritis …
Enbrel	TNF inhibitor	plaque psoriasis

Local Data Source

Relevant external entities

Generic Drugs

generic_name	adverse_effects
Adalimumab	After treatment with adalimumab …
Etanercept	Etanercept binds specifically to tumor …

External Data Source

Bio Compounds

formula	mechanisms
$C_{6428}H_{9912}N_{1694}O_{1987}S_{46}$	Binds with specificity to tumor …
$C_{2224}H_{3475}N_{621}O_{698}S_{36}$	There are two distinct receptors …

External Data Source

[Rare Daily Staff 2023] & [Ashburn & Thor 2004]

Augment Humira With that Relevant Info

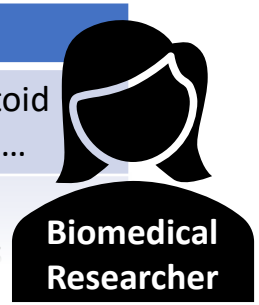
brand_name	class	uses	adverse_effects
Humira	TNF inhibitor	rheumatoid arthritis …	After treatment with adalimumab …

mechanisms
Binds with specificity to tumor …

FDA-Approved Drugs

brand_name	class	uses
Humira	TNF inhibitor	rheumatoid arthritis …
Enbrel	TNF inhibitor	plaque psoriasis

Local Data Source



Generic Drugs

generic_name	adverse_effects
Adalimumab	After treatment with adalimumab …
Etanercept	Etanercept binds specifically to tumor …

External Data Source

Bio Compounds

formula	mechanisms
$C_{6428}H_{9912}N_{1694}O_{1987}S_{46}$	Binds with specificity to tumor …
$C_{2224}H_{3475}N_{621}O_{698}S_{36}$	There are two distinct receptors …

External Data Source

[Rare Daily Staff 2023] & [Ashburn & Thor 2004]

Manually Querying for Relevant External Entities

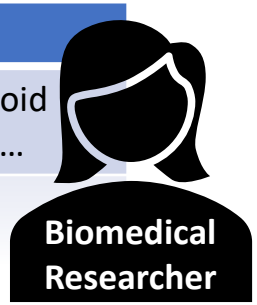
Challenges:

- Many external data sources
- Data heterogeneity: different representations
  - Humira = Adalimumab = $C_{6428}H_{9912}N_{1694}O_{1987}S_{46}$ = ???

FDA-Approved Drugs

brand_name	class	uses
Humira	TNF inhibitor	rheumatoid arthritis …
Enbrel	TNF inhibitor	plaque psoriasis

Local Data Source



Generic Drugs

generic_name	adverse_effects
Adalimumab	After treatment with
Et	

Search

External Data Source

Bio Compounds

formula	mechanisms
$C_{6428}H_{9912}N_{1694}O_{1987}S_{46}$	Binds with specificity
C_{2}	

Search

External Data Source

[Rare Daily Staff 2023] & [Ashburn & Thor 2004]

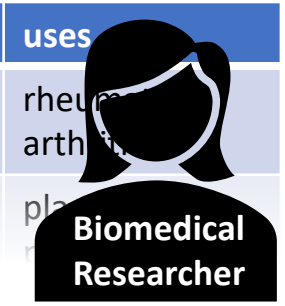
1st Try: Query = Too Specific to Local Source

Query = no relevant entities!

FDA-Approved Drugs

brand_name	class	uses
Humira	TNF inhibitor	rheu… arth…
Enbrel	TNF inhibitor	pla…

Local Data Source



Generic Drugs

generic_name	adverse_effects
Adalimumab	After treatment with

Humira | Search

0 results

External Data Source

Bio Compounds

formula	mechanisms
$C_{6428}H_{9912}N_{1694}O_{1987}S_{46}$	Binds with specificity

Humira | Search

0 results

External Data Source

[Rare Daily Staff 2023] & [Ashburn & Thor 2004]

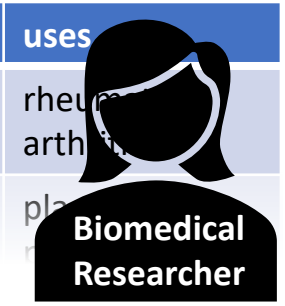
2nd Try: Query = Too General

Query = too many
non-relevant entities

FDA-Approved Drugs

brand_name	class	uses
Humira	TNF inhibitor	rheu… arth…
Enbrel	TNF inhibitor	pla…

Local Data Source



Generic Drugs

generic_name	adverse_effects
Adalimumab	After treatment with
Et…	

TNF plaque | Search

1090 results

External Data Source

Bio Compounds

formula	mechanisms
$C_{6428}H_{9912}N_{1694}O_{1987}S_{46}$	Binds with specificity
$C_{2…}$	

rheumatoid arthritis | Search

243 results

External Data Source

[Rare Daily Staff 2023] & [Ashburn & Thor 2004]

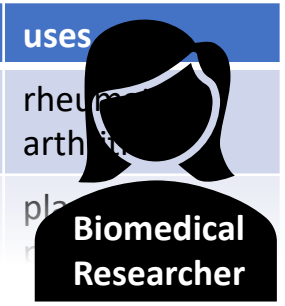
Nth Try: Just Right!

1. Retrieves relevant entity
2. …and few non-relevant entities

FDA-Approved Drugs

brand_name	class	uses
Humira	TNF inhibitor	rheu… arth…
Enbrel	TNF inhibitor	pla…

Local Data Source



Adalimumab

Generic Drugs

generic_name	adverse_effects
Adalimumab	After treatment with

TNF inhibitor crohns | Search

9 results

External Data Source

$C_{6428}H_{9912}N_{1694}O_{1987}S_{46}$

Bio Compounds

formula	mechanisms
$C_{6428}H_{9912}N_{1694}O_{1987}S_{46}$	Binds with specificity

psoriasis TNF arthritis | Search

14 results

External Data Source

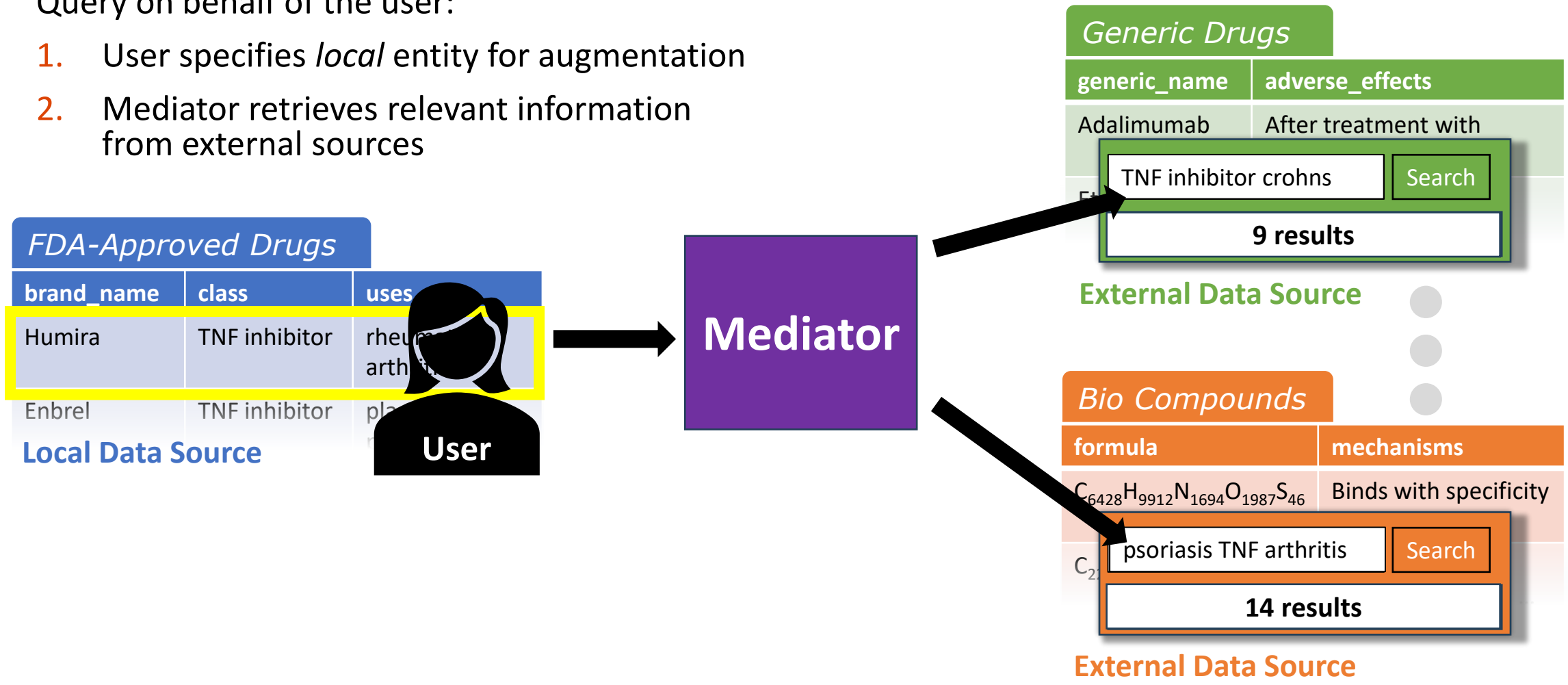
A lot of work!

[Rare Daily Staff 2023] & [Ashburn & Thor 2004]

Alternative: Use a Mediator

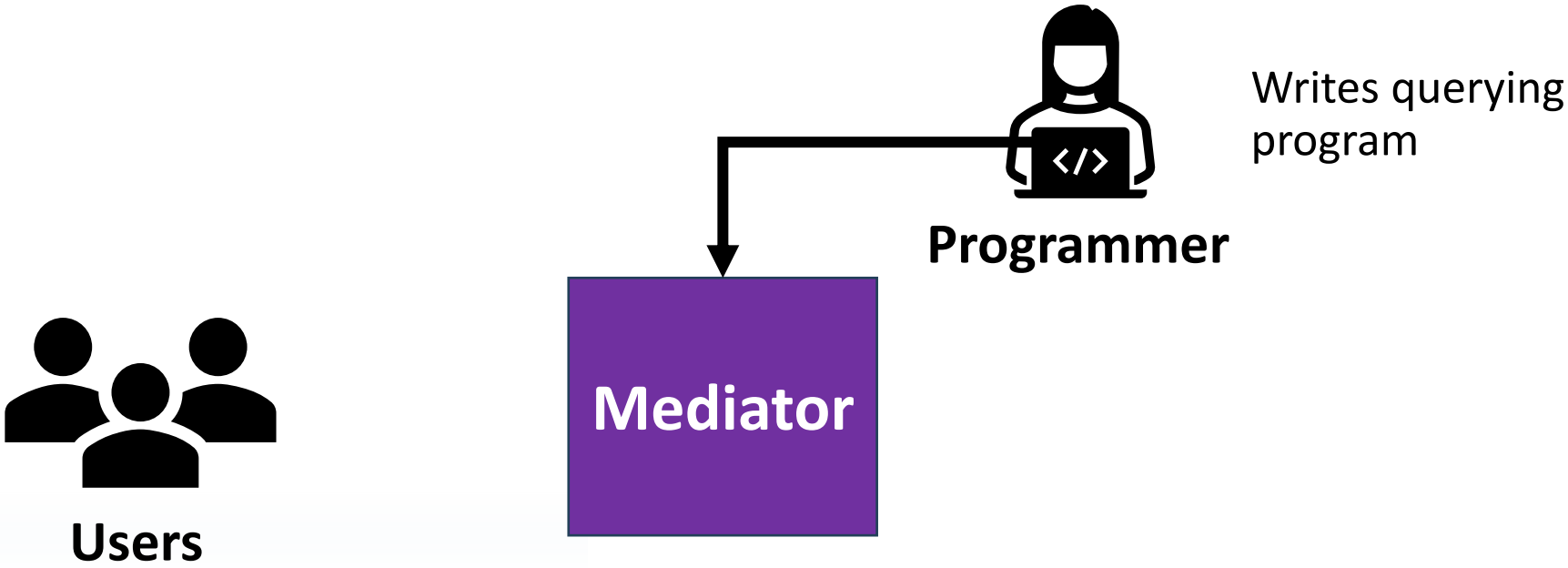
Query on behalf of the user:

1. User specifies *local* entity for augmentation
2. Mediator retrieves relevant information from external sources



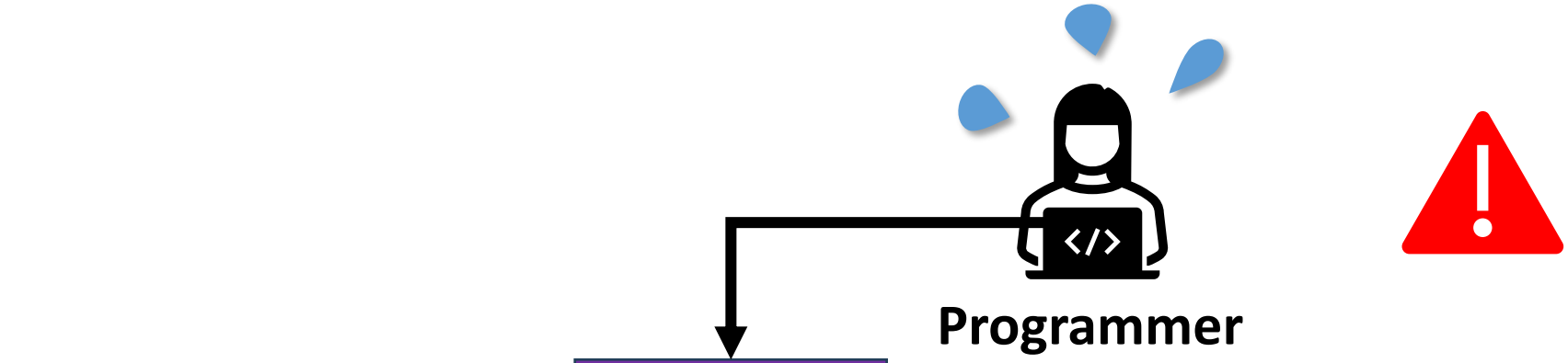
[Rare Daily Staff 2023] & [Ashburn & Thor 2004]

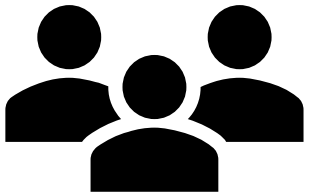
Existing Work: Mediator Written By Hand



[Rare Daily Staff 2023] & [Ashburn & Thor 2004]

Existing Work: Lots of Work!





Users

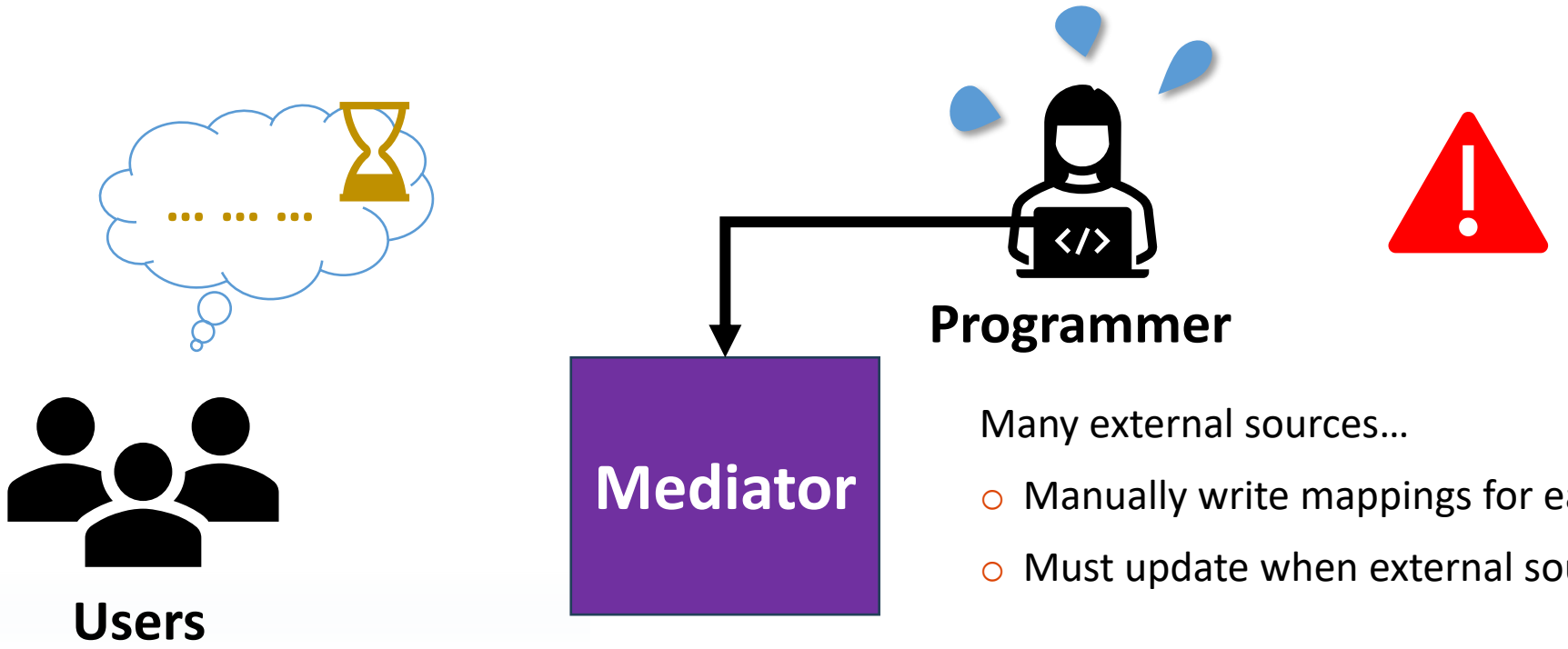
Mediator

Many external sources…

- Manually write mappings for each source
- Must update when external sources change

[Rare Daily Staff 2023] & [Ashburn & Thor 2004]

Existing Work: Information Delays

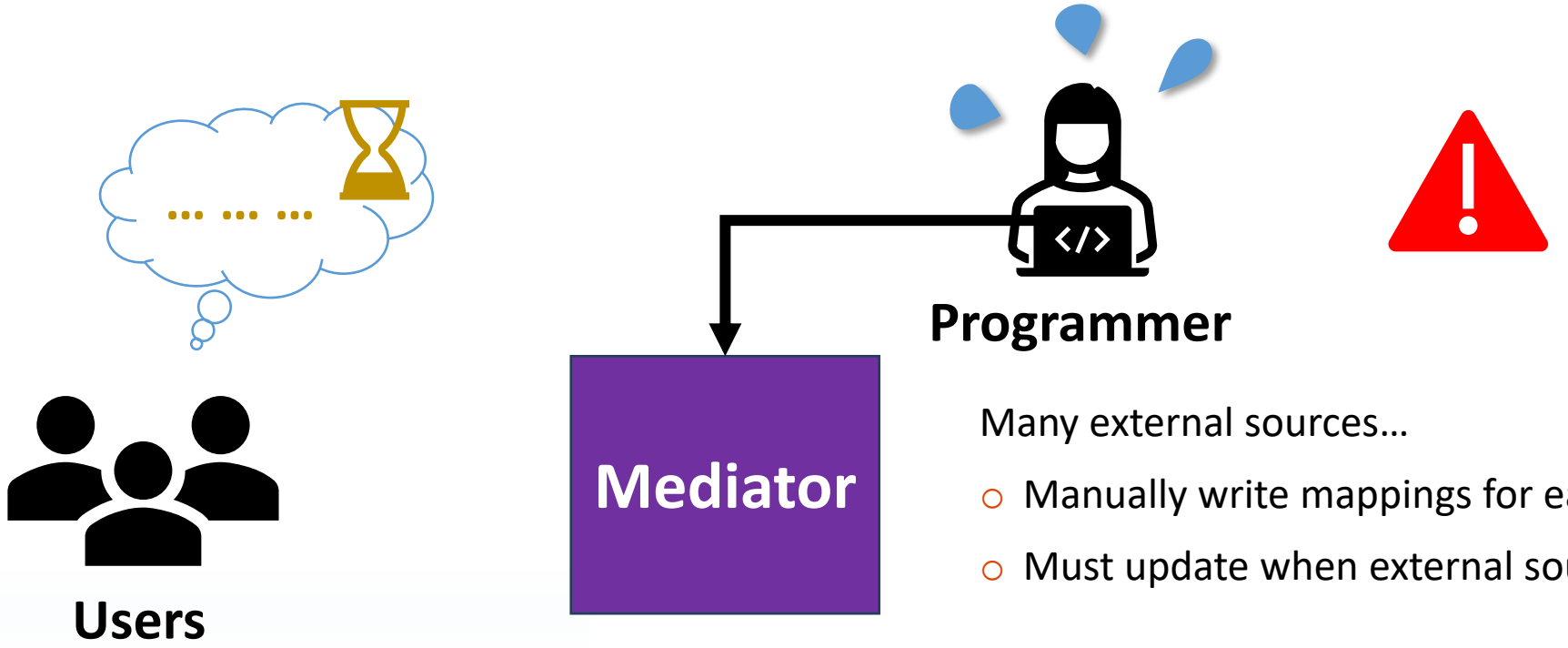


Many external sources…

- Manually write mappings for each source
- Must update when external sources change

[Rare Daily Staff 2023] & [Ashburn & Thor 2004]

Existing Work: Resource Intensive!



Many external sources…

- Manually write mappings for each source
- Must update when external sources change

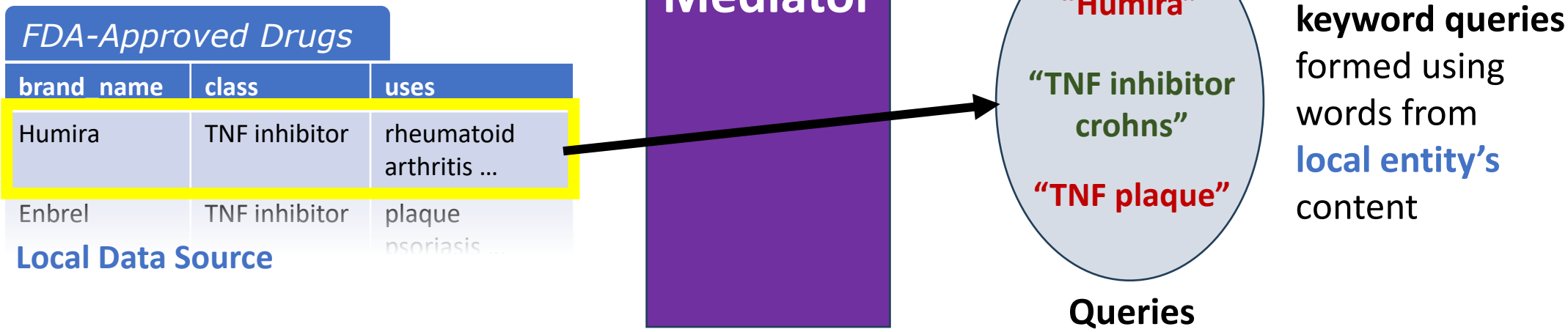
For example: the NIH funds a consortium of such systems (~14 systems)

- Just one system has 73 external datasources and millions of entities
- Costs NIH **US$923 million per year!**

Our Approach: Learn a Mediator

Reduce work and delays of writing the mediator *by hand*

Learn Mediator that maps **local entity** → **"Just right"** query



keyword queries formed using words from **local entity's** content

How Do We Learn the Mediator?

Offline Learning:

1. Gather training data
2. Train mediator
3. Users query mediator

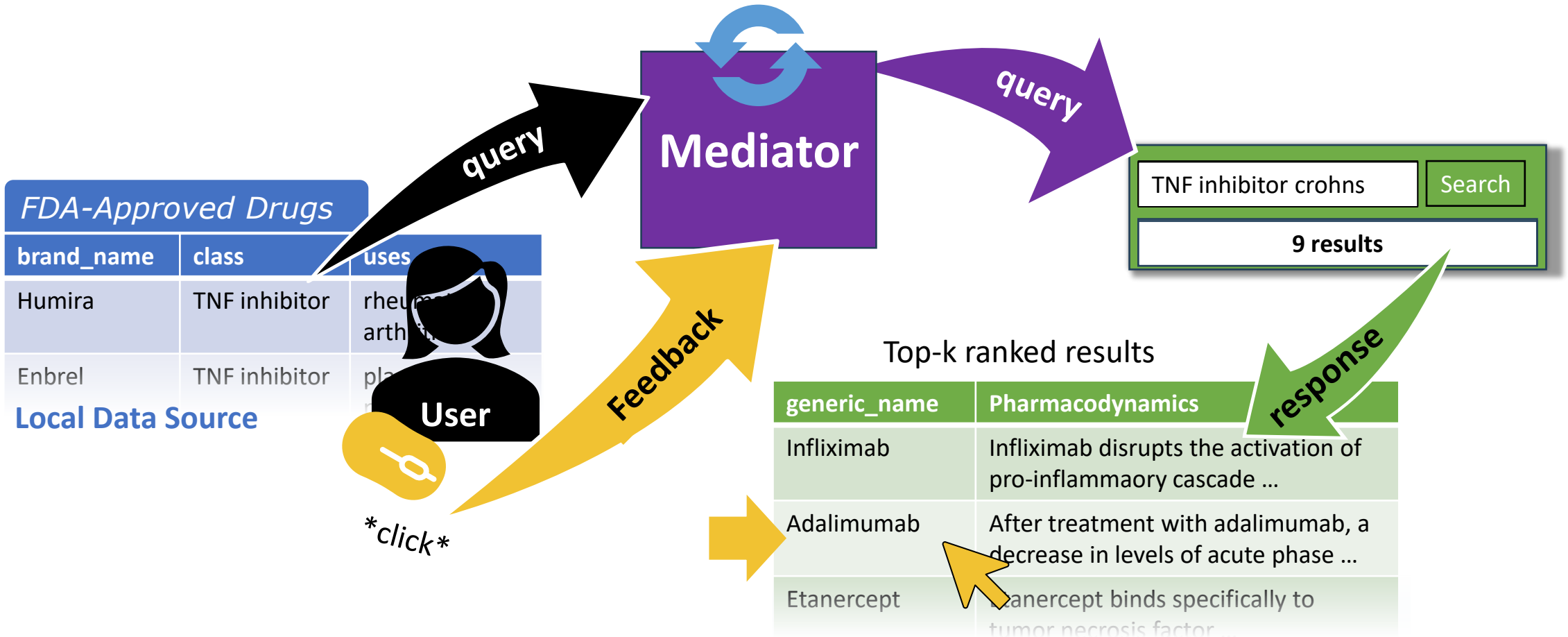
- Lots of expensive work
  - Hire domain experts to label data
  - External source updates → must repeat!
- Still delays…

Online Learning:

- Train mediator *while* users query it

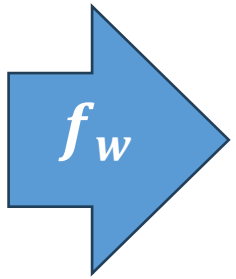
Online Learning Framework

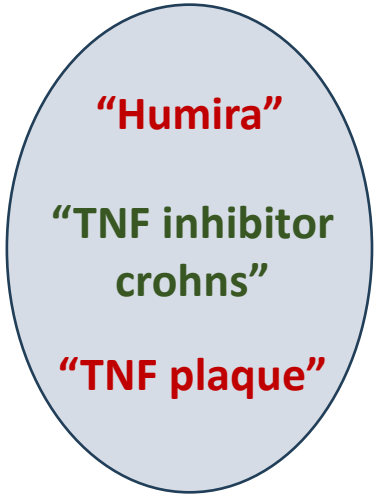
Refine understanding of what makes a query good



Predicting Query Quality with f_w

brand_name	class	uses
Humira	TNF inhibitor	rheumatoid arthritis …





Feedback is used to update w

Design Challenges:

Online setting: only know the quality of queries tried
Exploration: try new queries that *may* be better
Exploitation: use queries known to be good

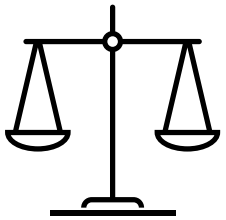
Short-Run Success: find sufficiently good queries quickly

- Users must remain engaged with the system

Dataset-Level: Fast and Lean

Idea: learn a simple predictor

$\mathbf{f_w}$ = linear function of **local entity's** features (*lexical*, *distributional*, and *schematic*)

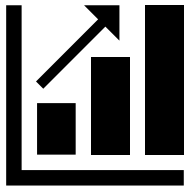


Encourage exploration in *feature-space*

Pro: will converge fast (simple function)

Con: not expressive; model may not work for every **local entity**

Design Challenge:



Long-Run Success: should *continue* to improve over time

- Methods should not waste feedback
- Avoid underfitting

Hybrid: Combine Simple Models

Idea: learn a set of simple models to combat underfitting

F = a set of linear models starting with $\{\mathbf{f_w}\}$ (Dataset-Level)

1. Converge

Domain	
f_w	$\{e_0, e_1, e_2, e_3\}$

Underfitting!

2. New models for hard entities

Domain	
f_w	$\{e_0, e_1, e_3\}$
$f_{w'}$	$\{e_2\}$

More feedback

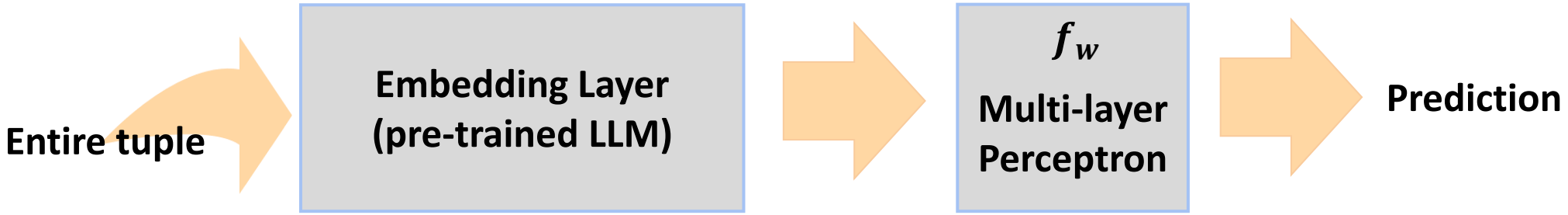
3. Increase overall performance

Domain	
f_w	$\{e_0, e_1, e_3\}$
$f_{w'}$	$\{e_2\}$

LMM: Using LLM Embeddings

Idea: leverage prior knowledge of a pre-trained large language model (LLaMA)

- Help in short-run and long-run



Every so often,
select a random
query instead of
the estimated best

Empirical Study Setup

Dataset	Source	Desc.	#entities
DrugCentral	Local	Molecular information specific to drugs	3,475
	External	Regulatory information about drugs	4,927
Drugs	Local	Drug reviews	13,725
	External	Wikipedia summaries of drugs	46,976
News	Local	Article titles and summaries	30,000
	External	Article content	30,000
WDC	Local	Products	57,109
	External	Products	55,247
ChEBI	Local	Molecular information specific to drugs	5,483
	External	Molecules and their effects on living organisms	189,467
CORD-19	Local	Abstract	250,575
	External	Title, authors, etc,.	340,826

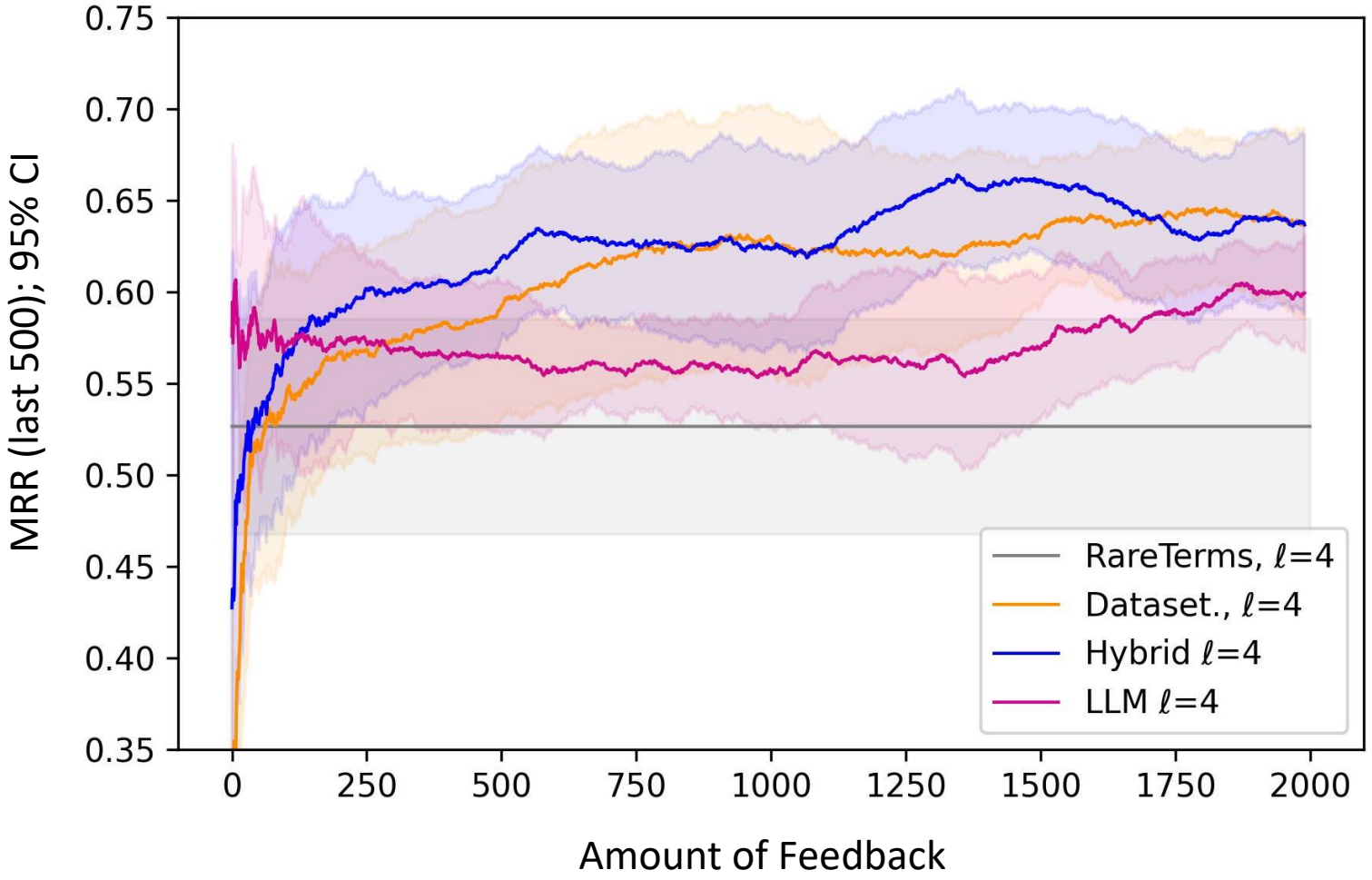
Run simulations over a wide variety of datasets

- Ground truth = feedback

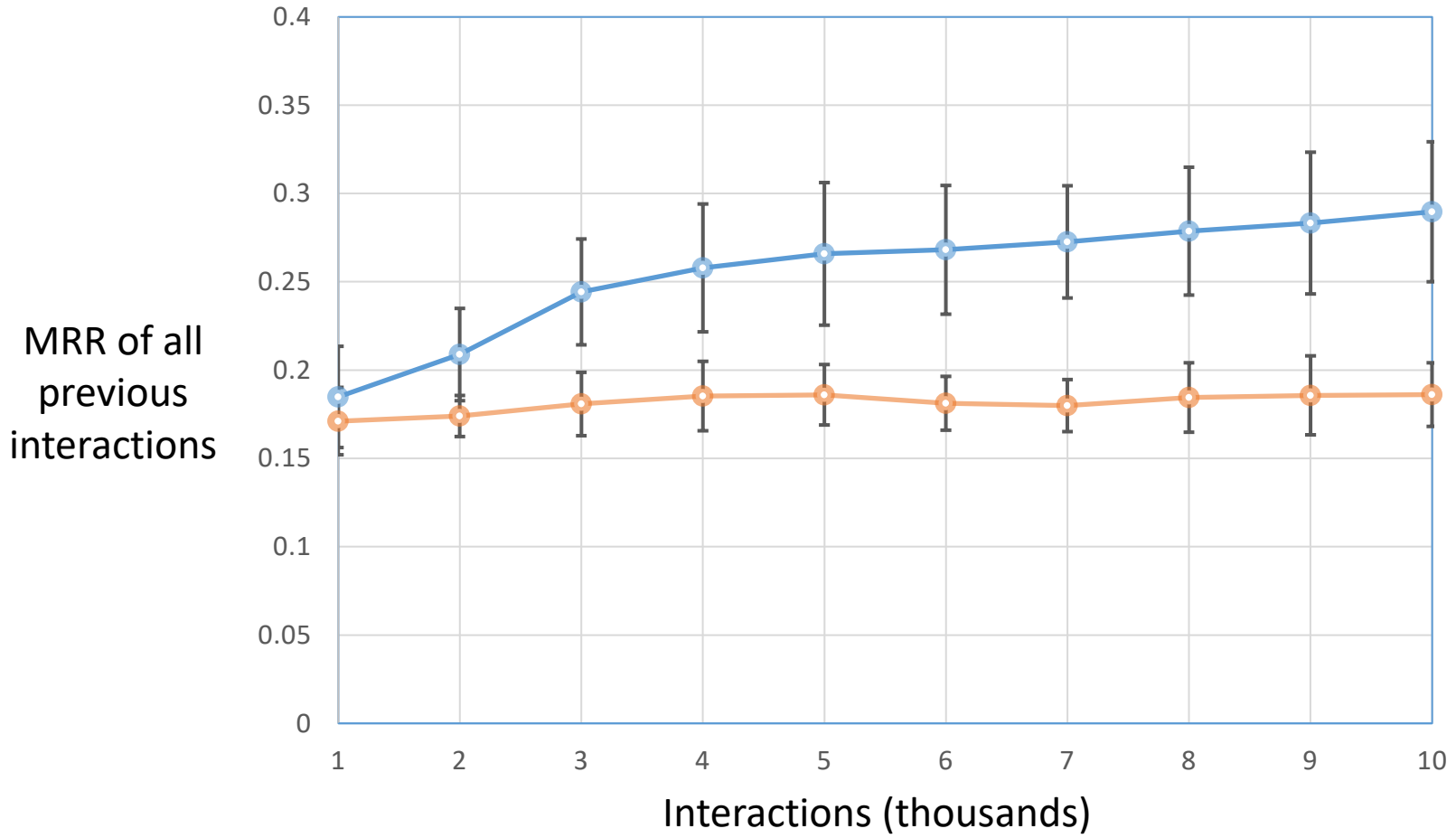
QUESTION: can our models…

- learn quickly?
- and keep learning?

Learning Results: ChEBI



Hybrid vs. Dataset-Level



Hybrid vs. **Dataset-Level**

- CORD-19
- Same stream of local entities

Takeaways

Motivation

- Mediators require a lot of resources to build/maintain by hand

Approach/Problem

- Learn the mediator online using user feedback!
- Methods to balance short-run and long-run success

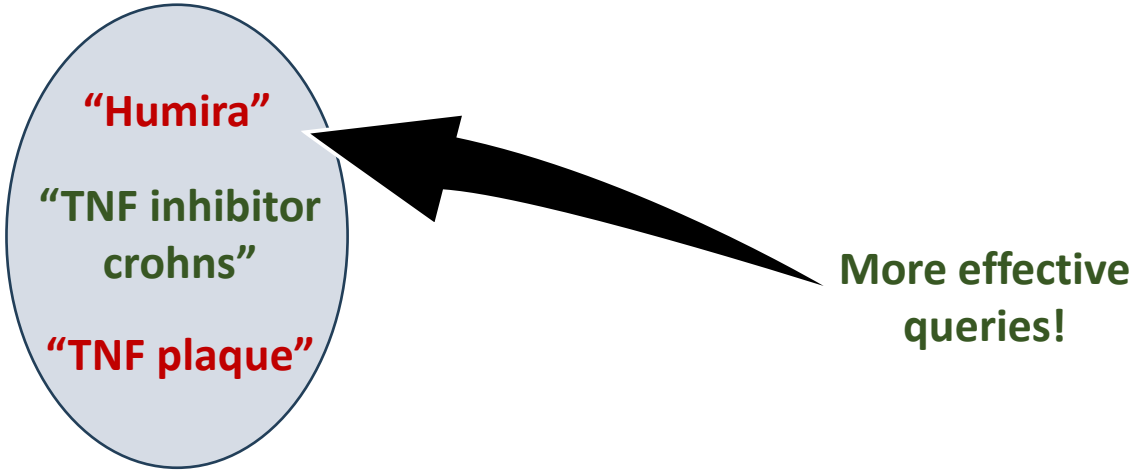
Experiments

- Experiments over a variety of data sets
- They do well!

Other Techniques (See our Paper!)

Term borrowing:

- Expand co-domain over time



Dynamic query length:

- Adjust number of terms in query automatically

Experiments using Longformer (another LLM)

…and more results!

Plug

Generating Data Augmentation Queries Using Large Language Models @ LLMDB 2023 (Friday)

Thank you!

Please share your questions!