IBM Research

QuerioDALI:
QA over Knowledge Graphs

Person-Centered Care

Vanessa Lopez, Pierpaolo Tommasi, Spyros Kotoulas, Jiewen Wu

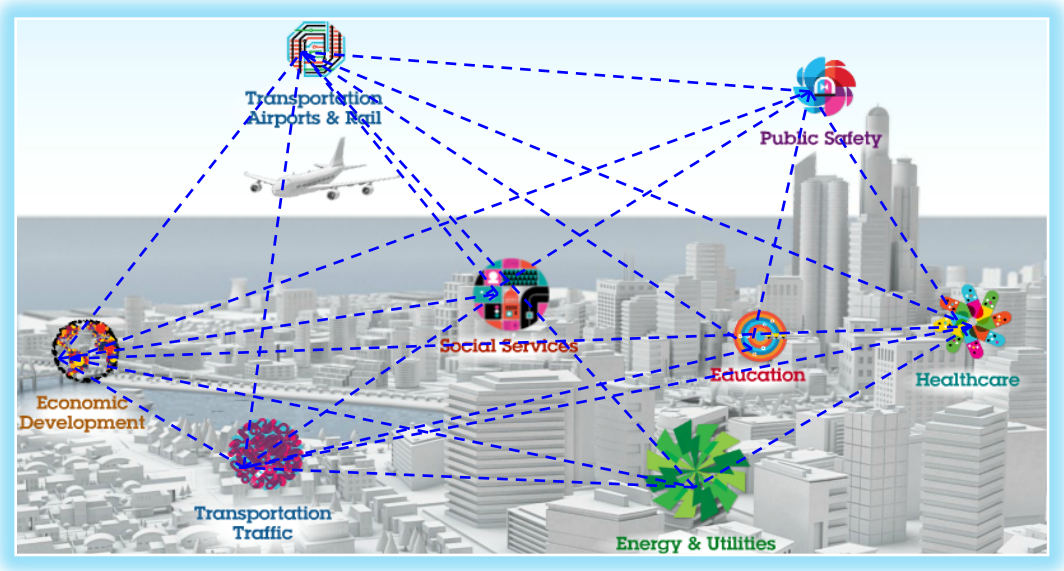
ISWC'16

Kobe, Japan

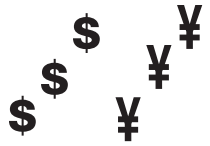
IBM

Goals

- Explore natural ways to **answer complex information needs** over distributed **knowledge graphs**, obtained from Linked Data sources, open and enterprise semi-structured data.



Use Case: Integrated Care



"It cost us one million dollars not to do something about Murray,"

Citizen-centric.

Social context.

DEPT. OF SOCIAL SERVICES

MILLION-DOLLAR MURRAY

Why problems like homelessness may be easier to solve than to manage.

BY MALCOLM GLADWELL

Murray Barr was a bear of a man, an ex-marine, six feet tall and heavy-set, and when he fell down—which he did nearly every day—it could take two or three grown men to pick him up. He had straight black hair and olive skin. On the street, they called him Smokey. He was missing most of his teeth. He had a wonderful smile. People loved Murray.

homeless people of Reno did, which is to walk through the casinos and finish off the half-empty glasses of liquor left at the gaming tables.

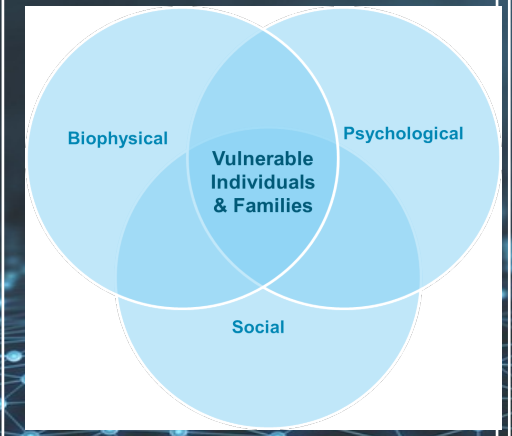
"If he was on a runner, we could pick him up several times a day," Patrick O'Bryan, who is a bicycle cop in downtown Reno, said. "And he's gone on some amazing runners. He would get 'Murray, you know you love us,' and he'd say, 'I know'—and go back to swearing at us."

"I've been a police officer for fifteen years," O'Bryan's partner, Steve Johns, said. "I picked up Murray my whole career. Literally."

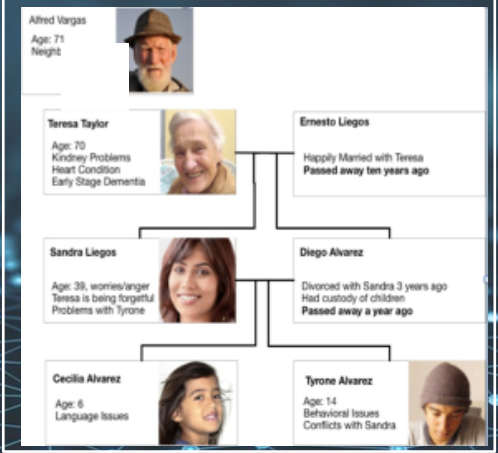
Johns and O'Bryan pleaded with Murray to quit drinking. A few years ago, he was assigned to a treatment program in which he was under the equivalent of house arrest, and he thrived. He got a job and worked hard. But then the program ended. "Once he graduated out, he had no one to report to, and he needed that," O'Bryan said. "I don't know whether it was his military background. I suspect that it was. He was a good cook. One time, he accumulated

A 360° person view

Citizen



Social Network



Care Network

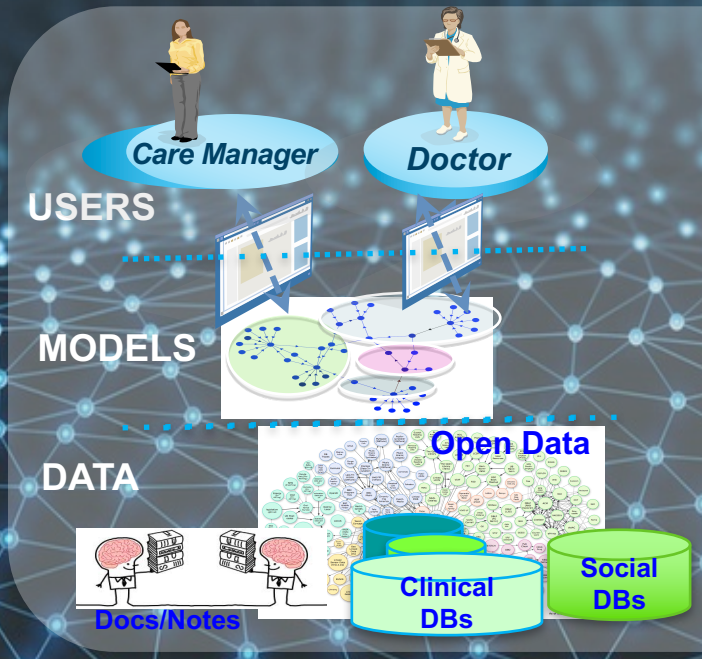
Care Team

Transportation

Time

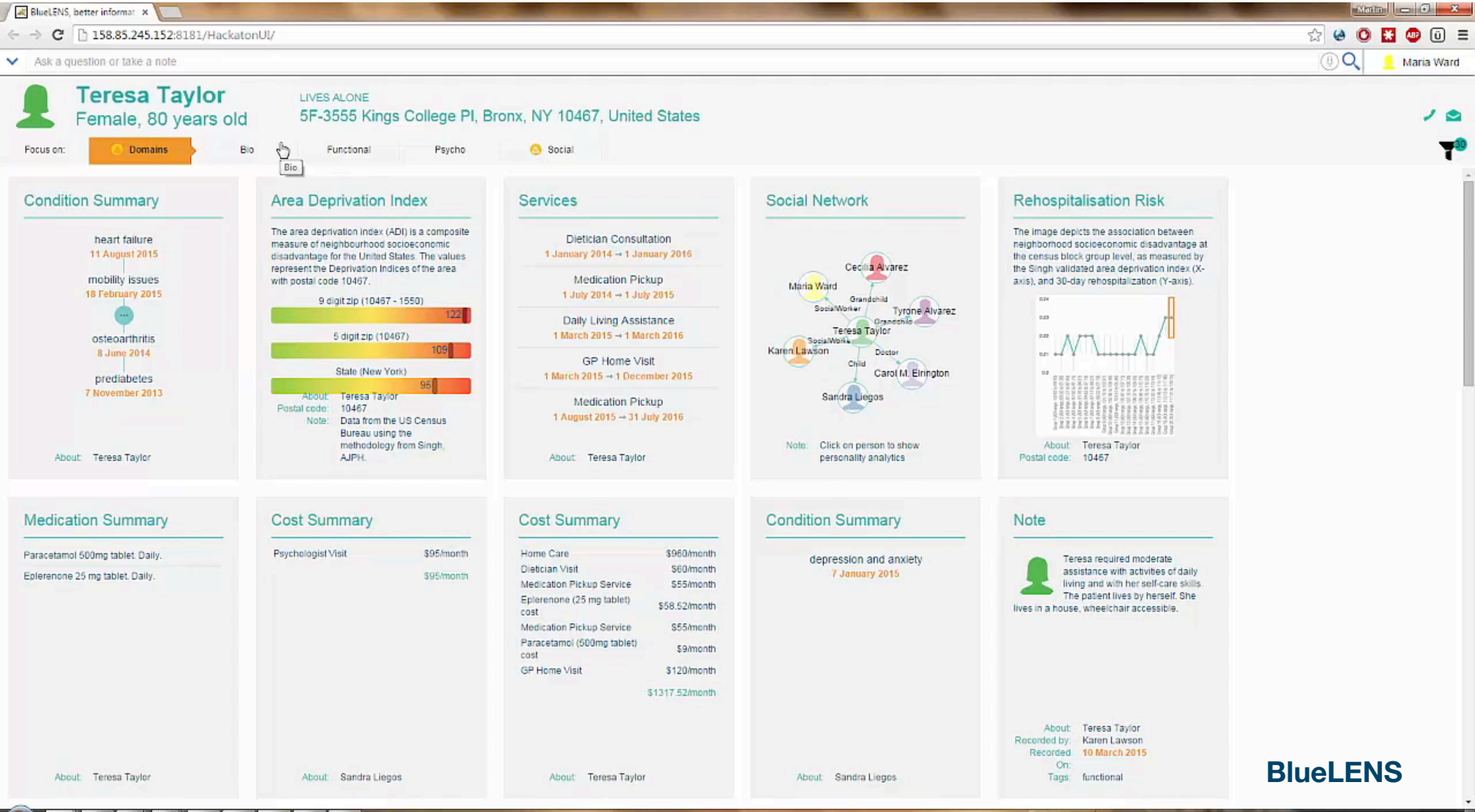
A 360◦ person view

- **Semantics** for understanding and interlinking complex data
- **Contextual** what is relevant depends on the user task
- **NL** for users to **interact** with the system



Research Agenda

- How can we make this information available to a care team in a natural way?
  Right info at right time

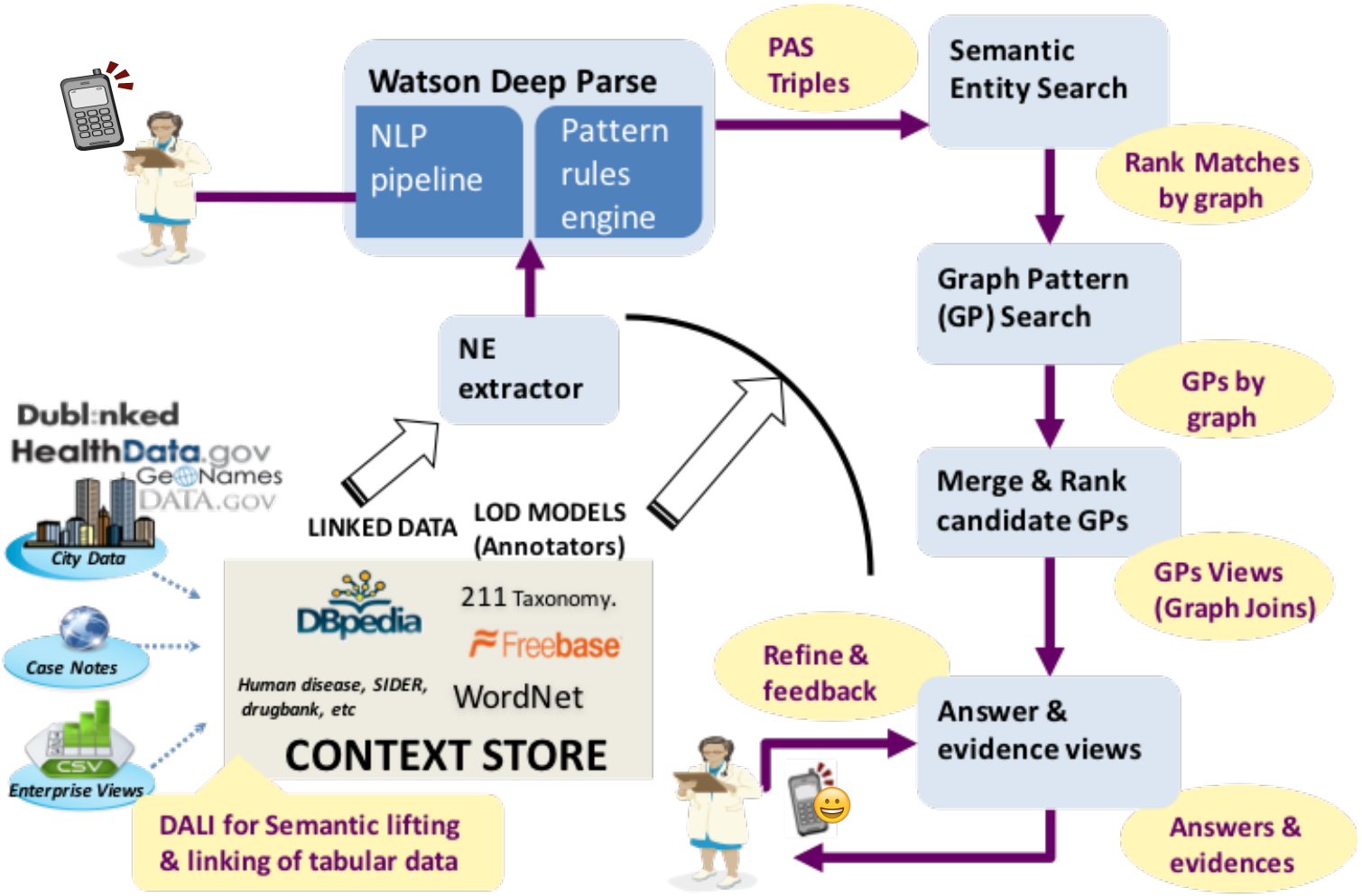


BlueLENS

Requirements for QA

- Open Domain
- No fixed schema
- Dynamic and evolving
- No training data
- Combining facts across sources

Approach: QA Pipeline



BlueMix: https://console.ng.bluemix.net/catalog/

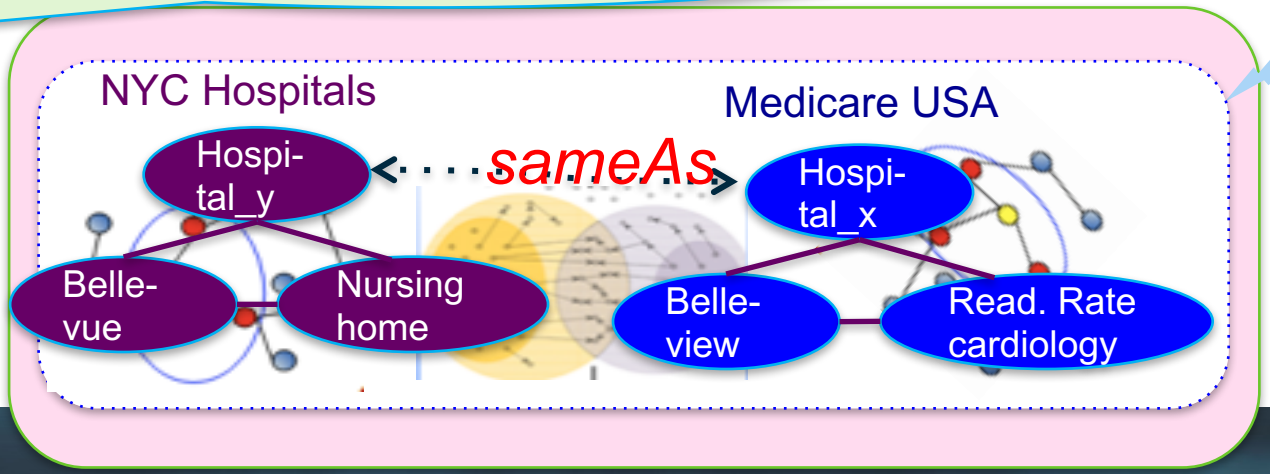
Linked Repositories & Annotators

Which hospitals with elderly care have the lowest readmission rates for cardiology in NYC?

Context

Enterp. Data

Open Data



Views

Freebase **1.9 billion** RDF triples

SIDER

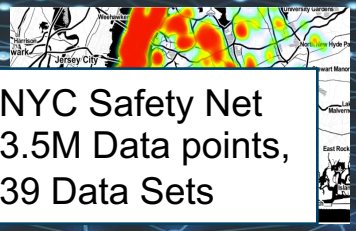
12 million Concept names in UMLS

DRUGBANK

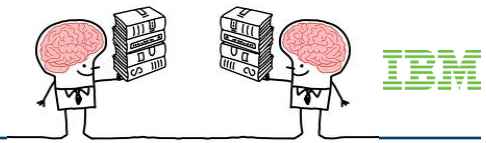
580 million RDF triples in English DBPedia

Diseasome

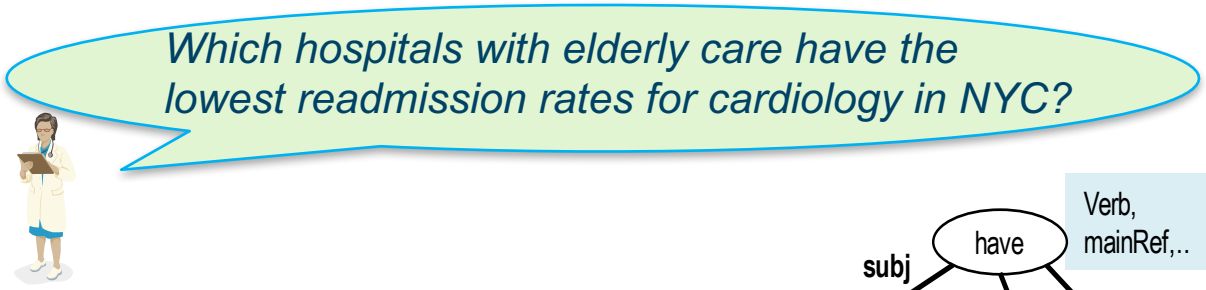
WordNet



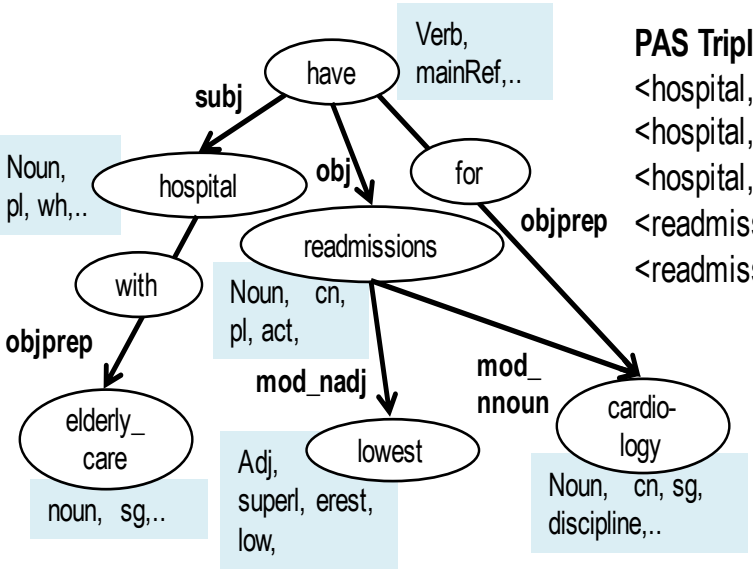
NYC Safety Net
3.5M Data points,
39 Data Sets



Step 1: NL Deep Parsing



- From dependency tree to PAS triples
  - Watson deep parsing (UIMA)
- NE extraction (multi-words: elderly care)

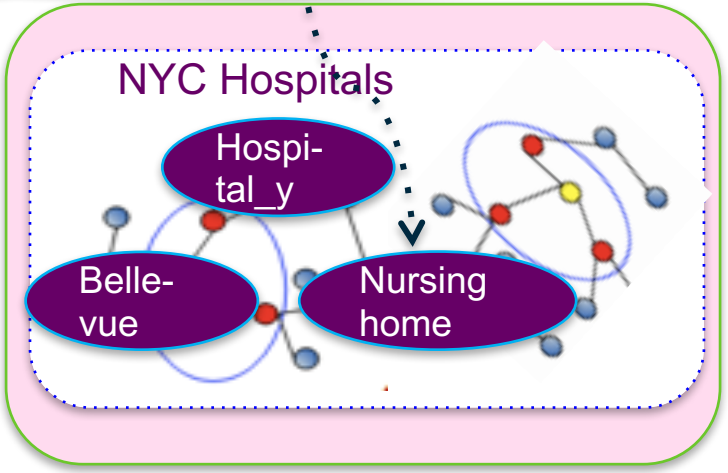


PAS Triples:
<hospital, have, cardiology>
<hospital, elderly care>
<hospital, have, readmission>
<readmission, cardiology>
<readmission, low>

pattern=nounAdjSup ->
nsubjVar[hasPartOfSpeech("noun")]
{ mod_nadj|mod_nadv ->
ndetVar[hasParseFeature("superl")]
}

Step 2: Semantic Entity Search

*Which **hospitals** with **elderly care** have the lowest readmission rates for cardiology in NYC?*



- Candidate URIs for each query term
  - Mix of schema / non-schema: dateOfbirth, birthDate, birthdate, ..
  - Duplicated terms / missing types: spouse, husband, married to, ..
- Balancing P/R:
  - Semantic expansion
  - Confidence score

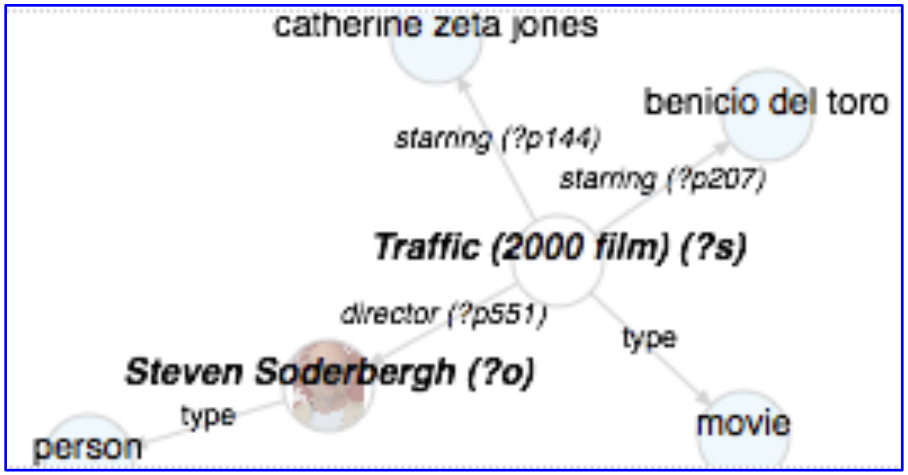
Step 3: Graph Pattern Search

- Candidate entities in the PAS are map (partially or completely) into *Graph Patterns*, which are combined into *views* that answer the query
- Large space of mapping combinations:
  - Find most precise (coverage, less expensive) interpretations first

Question: Who is the director of movies starring Benicio del Toro and Catherine Zeta-Jones?

PAS Triples: <person/org., direct, movies>
<movies, starring, Benicio del Toro>
<movies, starring, Catherine Zeta-J>

Focus: person/org
Join Term: movies
Answer: Steven Soderberg (movie: traffic)



A GP can be translated it into SPARQL and consists of:
- BGPs that belong to the same graph
- JOINS AND UNIONS between the BGPs
- Solution modifiers such as as : ORDER BY DESC/ASC , COUNT, OFFSET and LIMIT
- A confidence score
- The variables that are the focus of the GP

GP Templates and Semantic Relatedness

IBM Research Querio Answering

Which mountains are higher than the Nanga Parbat?

Answer (score: 0.69, percent: 100)

- Kangchenjunga, 8586
- K2, 8611
- Mount_Everest, 8848
- Cho_Oyu, 8201
- Lhotse, 8516
- Makalu, 8481
- Dhaulagiri, 8167
- Manaslu, 8156
- Banua_Wuhu, –5

http://dbpedia.org

Boolean Comparative Pattern

Find hidden connections on the fly

dbpedia:Elevation

↓ dcterms:subject

Category:Length

↑ dcterms:subject

dbpedia:Height

↑ dbpedia:disambiguates

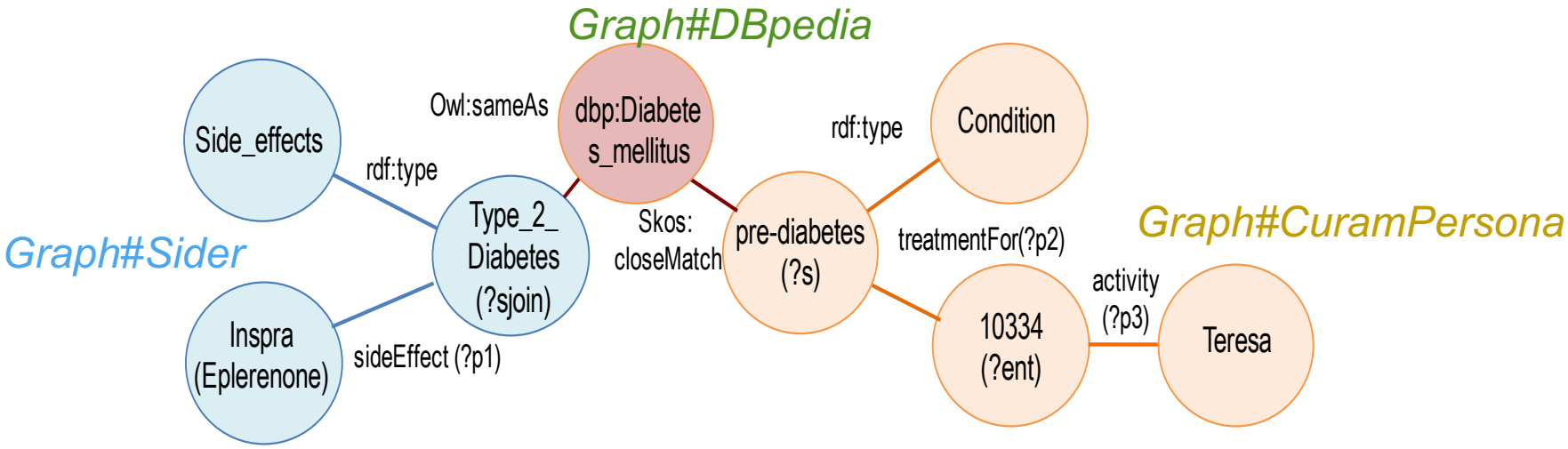
dbpedia:High

```
Select distinct ?s ?o1 where { {
?s rdf:type <dbo:Mountain> . ?s ?property ?o1.
 <dbr:Nanga_Parbat> ?property ?o2.
FILTER ( ?o1 > ?o2). FILTER (?property = <dbp:elevationM> )} }
```

Step 4: Merging & Ranking

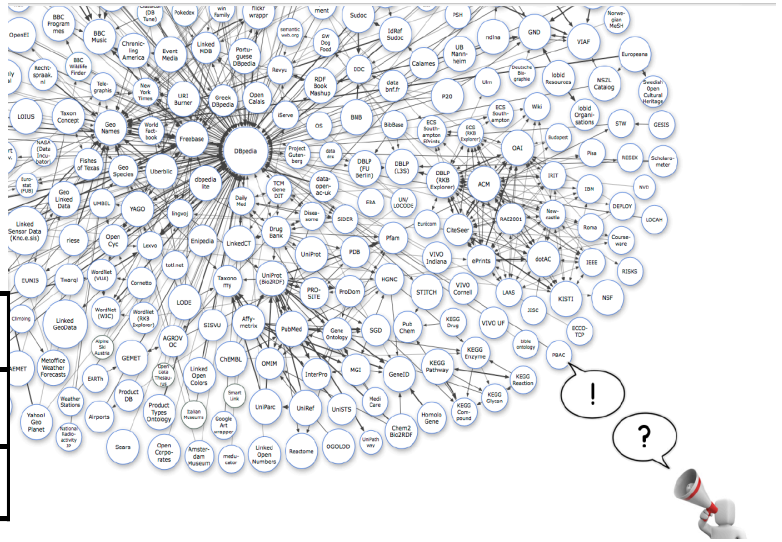
- Merging facts and partial answers across sources
- Entity co-reference (join term): different URIs for same *world* entity
  - Syntactic (similar labels) & Semantic Merging (entity linkage)
- Rank alternative translations

Question: Is Eplerenone having side effects for Teresa's conditions?
Join terms: "side effects" in SIDER ontology ("the side effects of Eplerenone") and conditions in the patient data ("conditions of Teresa")
Answer: yes



QALD & Free917 Benchmarks

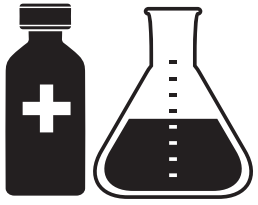
- Prototype 1: DBpedia (triples: 204,954,931)
  - QALD-5 50 questions
- Prototype 2: Freebase (triples: ????)
  - Free917 101 questions



	P@1	R@1	P@2	R@2	P@3	R@3	F-1	F-2	F-3
QALD5	**0.64**	**0.69**	0.53	0.73	0.52	0.73	**0.61**	0.55	0.55
Free917	**0.72**	**0.72**	0.63	0.73	0.62	0.73	**0.72**	0.65	0.64

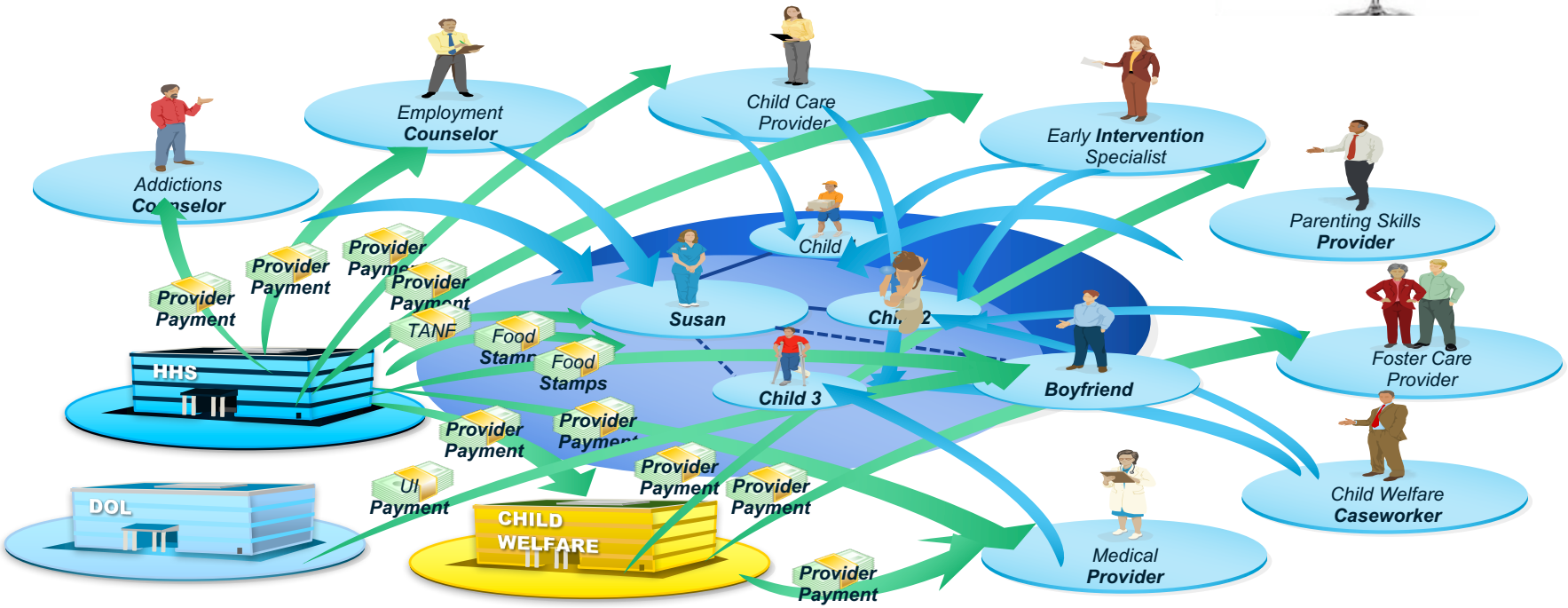
- Prototype 3: Biomedical Linked Open Data
  - QALD-4 25 questions
  - Sider (number of triples: 101,558)
  - Drugbank (number of triples: 517,144)
  - Diseasome (number of triples: 92,079)

P@1	R@1	P@2	R@2	P@3	R@3	F-1	F-2	F-3
0.85	0.88	0.78	0.92	0.78	0.92	0.85	0.83	0.83

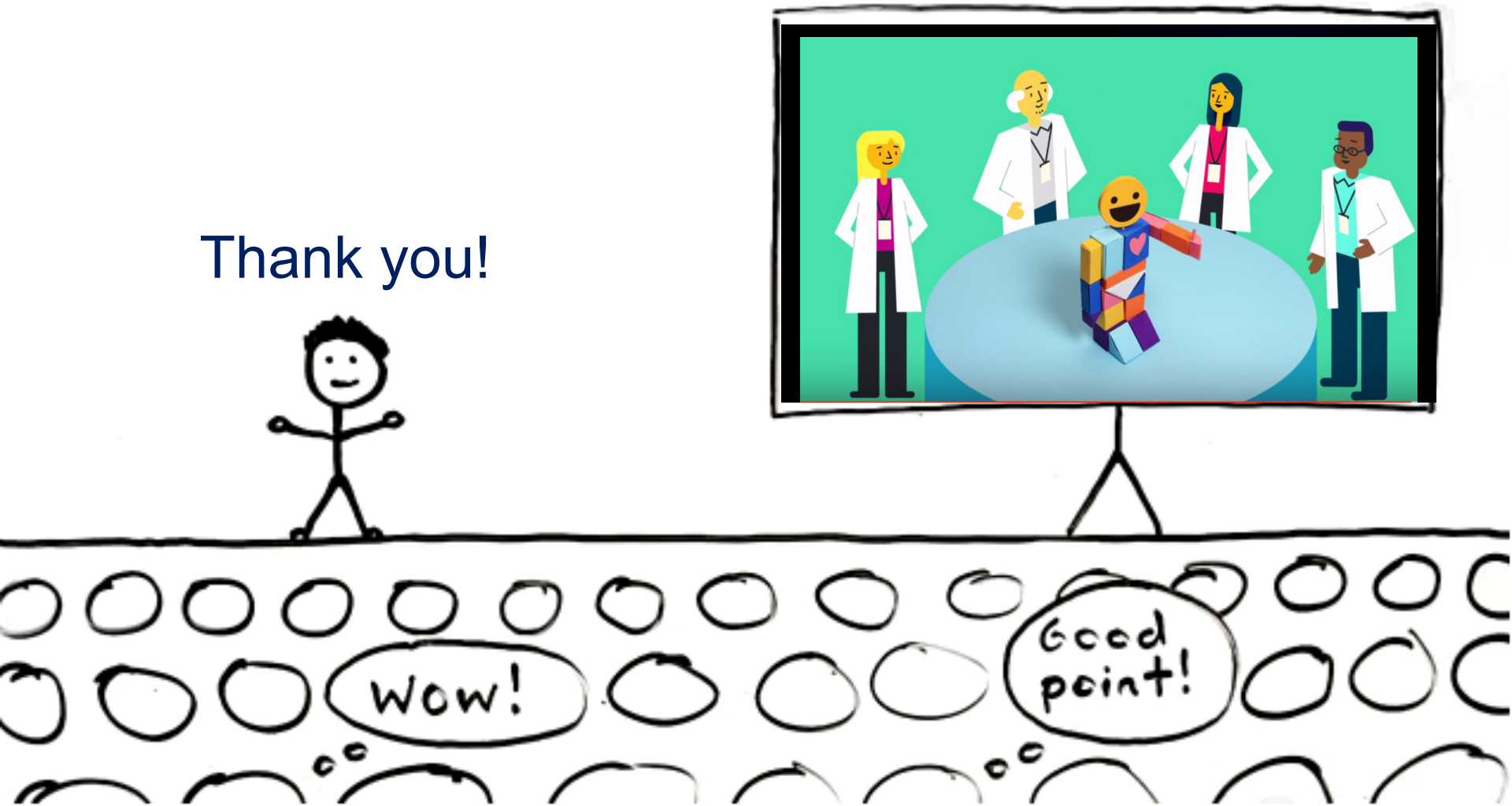


Many Open Challenges

- Open Domain QA is very hard
- Brittleness and scalability
- Hybrid QA
- More cognitive: Active learning from user interactions
- Personalization: roles



Thank you!



Evaluation

- Errors led to noisy answers (precision) or missing answers (recall)
- Linguistic: inaccurate parse tree or not enough rule coverage (e.g., what is the height difference between Mount Everest and K2?)
- Anchoring: not finding or filtering out the right mapping or multi-word term (e.g., democrat to “Democratic _party (United States))
- Patterns coverage: to bridge the lexical and structural gap (e.g., queries with temporal reasoning – since 2000 – indirect - *who was born in Halloween?)*
- Merging*:* failed to do the linking across join terms
- Ranking: not enough context (e.g., popularity)

Graph Pattern Search: Templates

Pattern1: E.g.: Is anorexia an eating disorder?
Direct relation entity to entity (Focus: ?p)
<entity1> ?p <entity2>
UNION <entity2> ?p <entity1>

Boolean (is-a) query (yes/no)

Pattern2: E.g.: rivers in Russia
Direct instance to type (Focus: ?s)
?s ?property <type>. {?s ?px <instance>}
UNION {<instance> ?px s}

Factoid (answer: list of entities)

Pattern 4: E.g.: population in countries
Direct type to property (Focus: ?s ?o)
?s ?property <type>. {?s <prop> ?o}
UNION {?o <prop> ?s>}

Factoid (answer: entities/*datatypes*)

Pattern 6: E.g.: highest mountain; who produced the most films
Superlative pattern (Focus: ?o)
Pattern 4+ order by desc(?o) offset 1 limit 0
Pattern4/5+ order by desc(count(?o)) offset 0 limit 1

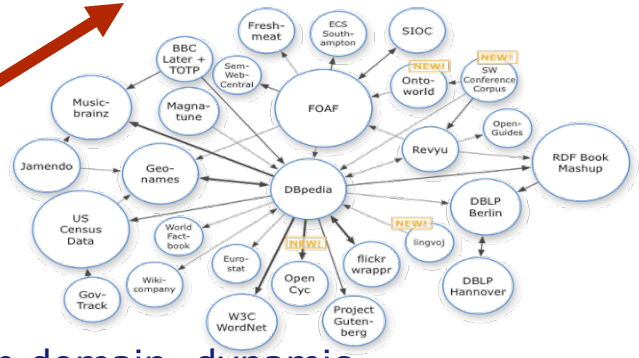
Factoid (superlative)

Pattern 7: E.g.: Is Lake Baikal bigger than the Great Bear Lake?
Comparative pattern Instance to Instance
<instance1> <prop> ?o1. <instance2> <prop> ?o2.
FILTER (?o1 > ?o2)

Query : Which artists were born on the same date as Rachel Stevens? (JOIN: date – property-)
SPARQL: select distinct ?s ?o where { {
?s ?property <http://dbpedia.org/ontology/Artist> .
?s <http://dbpedia.org/property/birthDate> ?o.
<http://dbpedia.org/resource/Rachel_Stevens> <http://dbpedia.org/property/birthDate> ?o. } }

QA over KGs: Scope

Open Semantic Web content



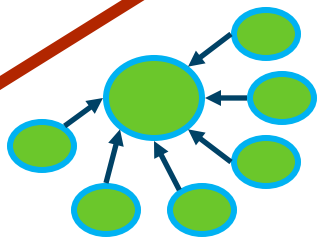
Global-view (proprietary) KG

Evi, Wolfram Alpha, Facebook KG, Google KG

- Open domain, dynamic
- Heterogeneous and distributed
- Large scale, noisy data
- Varying levels of quality and trust

Domain-specific

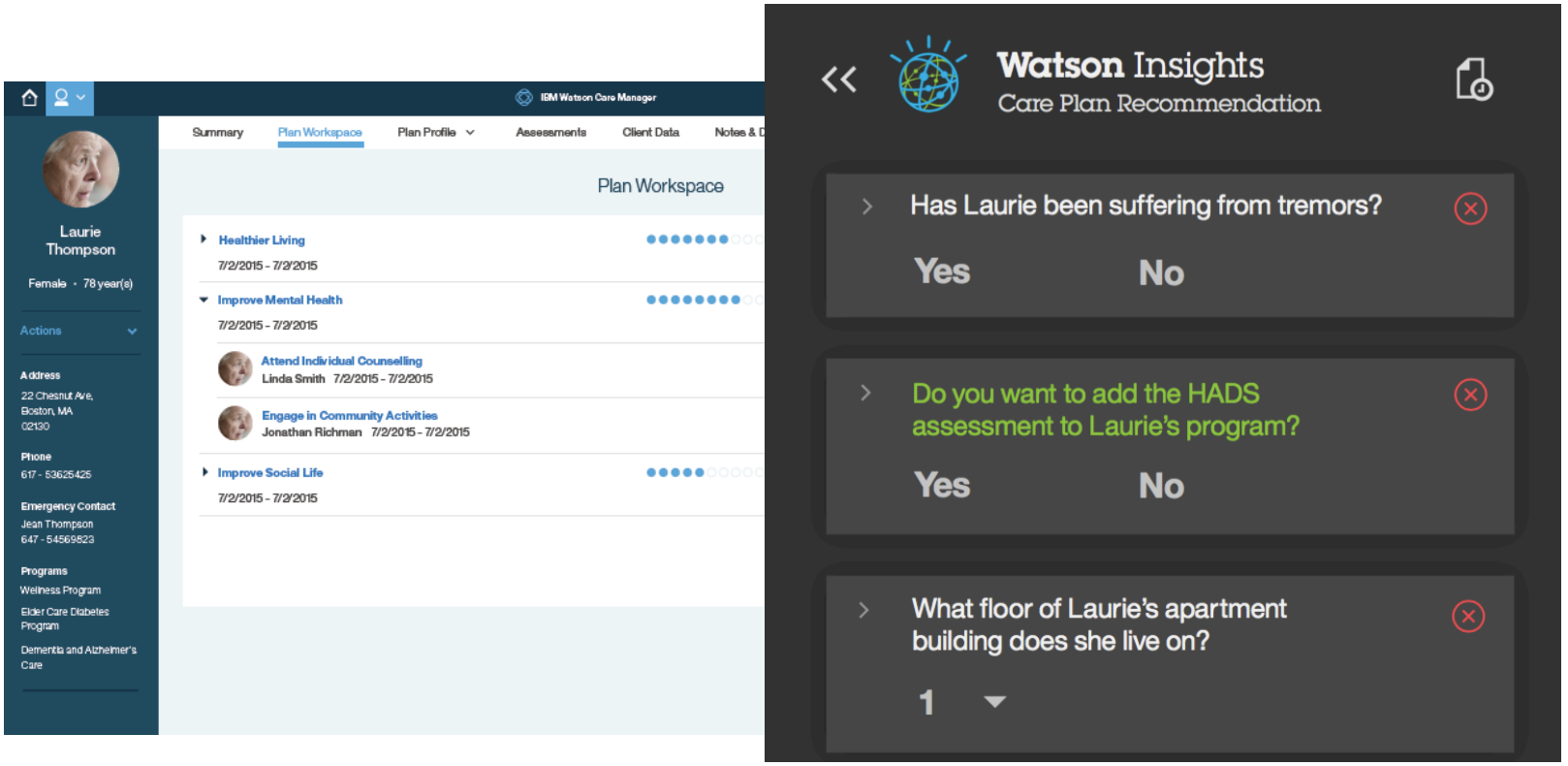
ontology-based QA, NLIDBs



- Open domain, dynamic
- Build their own large scale KG
- Homogeneous trusted

DESIGN CRITERIA:

-Domain-specific
-Portability (new twist on traditional NLDB issues)
-Medium-sized homogeneous Knowledge Base

Research Agenda

- How can we make this information available to a care team in a natural way? **Right info at right time**
- How to capture knowledge and best practices from care workers to provide **actionable insights**?