Managing General and Individual Knowledge in Crowd Mining Applications

Yael Amsterdamer, Susan Davidson, Anna Kukliansky,
Tova Milo, Slava Novgorodov and Amit Somech

CIDR 2015

Motivation

Ann, a vacationer, is interested in finding child-friendly activities at an attraction in NYC, and a good restaurant nearby (plus relevant advice).

Motivation

Ann, a vacationer, is interested in finding child-friendly activities at an attraction in NYC, and a good restaurant nearby (plus relevant advice).

"You can play baseball in Central Park and eat at Maoz Vegetarian.
Tips: Apply for a ballfield permit online"

"You can go visit the Bronx Zoo and eat at Pine Restaurant.
Tips: Order antipasti at Pine.
Skip dessert and go for ice cream across the street"

Motivation

Ann, a vacationer, is interested in finding child-friendly activities at an attraction in NYC, and a good restaurant nearby (plus relevant advice).

“You can play baseball in Central Park and eat at Maoz Vegetarian.
Tips: Apply for a ballfield permit online”

“You can go visit the Bronx Zoo and eat at Pine Restaurant.
Tips: Order antipasti at Pine.
Skip dessert and go for ice cream across the street”

Motivation

Ann, a vacationer, is interested in finding child-friendly activities at an attraction in NYC, and a good restaurant nearby (plus relevant advice).

"You can play baseball in Central Park and eat at Maoz Vegetarian.
Tips: Apply for a ballfield permit online"

"You can go visit the Bronx Zoo and eat at Pine Restaurant.
Tips: Order antipasti at Pine.
Skip dessert and go for ice cream across the street"

Motivation

Ann, a vacationer, is interested in finding child-friendly activities at an attraction in NYC, and a good restaurant nearby (plus relevant advice).

"You can play baseball in Central Park and eat at Maoz Vegetarian.
Tips: Apply for a ballfield permit online"

"You can go visit the Bronx Zoo and eat at Pine Restaurant.
Tips: Order antipasti at Pine.
Skip dessert and go for ice cream across the street"

A dietician may wish to study the culinary preferences in some population, focusing on food dishes that are rich in fiber

Motivation

Ann, a vacationer, is interested in finding child-friendly activities at an attraction in NYC, and a good restaurant nearby (plus relevant advice).

"You can play baseball in Central Park and eat at Maoz Vegetarian.
Tips: Apply for a ballfield permit online"

"You can go visit the Bronx Zoo and eat at Pine Restaurant.
Tips: Order antipasti at Pine.
Skip dessert and go for ice cream across the street"

A dietician may wish to study the culinary preferences in some population, focusing on food dishes that are rich in fiber

Motivation

Ann, a vacationer, is interested in finding child-friendly activities at an attraction in NYC, and a good restaurant nearby (plus relevant advice).

“You can play baseball in Central Park and eat at Maoz Vegetarian.

General knowledge:

- General truth, objective data, not associated with an individual
- *E.g., geographical locations*
- Can be found in a knowledge base or an ontology

Individual knowledge:

- Related to the habits and opinions of an individual
- *E.g., travel recommendations*
- We can ask people about it

A dietician may wish to study the culinary preferences in some population, focusing on food dishes that are rich in fiber

Motivation

Ann, a vacationer, is interested in finding child-friendly activities at an attraction in NYC, and a good restaurant nearby (plus relevant advice).

“You can play baseball in Central Park and eat at Maoz Vegetarian.

General knowledge:

- General truth, objective data, not associated with an individual
- *E.g., geographical locations*
- Can be found in a knowledge base or an ontology

When missing in the knowledge base, we can ask the crowd!

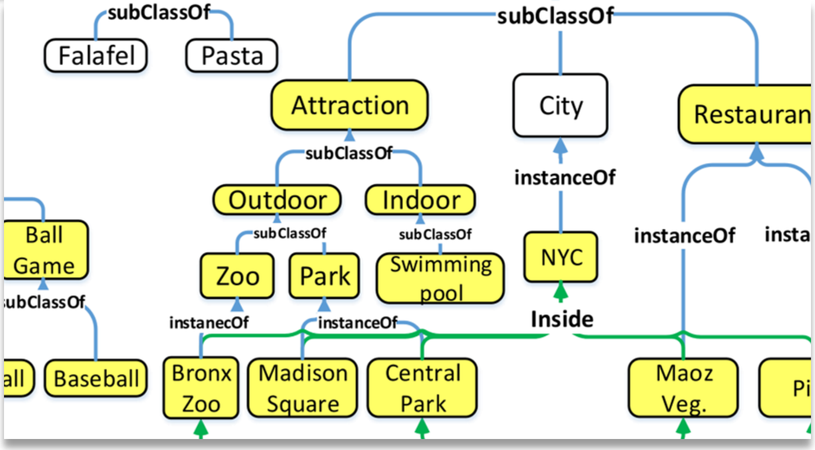
Individual knowledge:

- Related to the habits and opinions of an individual
- *E.g., travel recommendations*
- We can ask people about it

Crowd answers can be recoded in a knowledge base

Crowd Mining: Crowdsourcing in an Open World

Given an ontology of general knowledge and a mining task



- Incrementally explore relevant patterns

```
{Ball_Game playAt Central_Park}
```

- Generate (closed and open) questions to the crowd about them

*How often do you **play ball games** at **Central Park**?*

*Which **ball games** do you **play** at **Central Park**?*
*What else do you do at **Central Park**?*

- Evaluate the significance of the patterns and discover related ones

Pattern score = 0.6

```
{Baseball playAt Central_Park.
 Permit   getAt  "www.permits.org"}
```

- Produce a concise output that summarizes the findings

Crowd Mining Framework Design

We design a general architecture which outlines the components of a crowd mining framework and the interaction between them

Challenges:

Compiling user requests into a declarative query language

Deciding which questions to generate to the crowd next

How to aggregate crowd answers?

Personalization and crowd member selection

Updating and managing the knowledge base

Combining the crowd answers with knowledge base data

The type of processed data (general versus individual) must be taken into account

Today

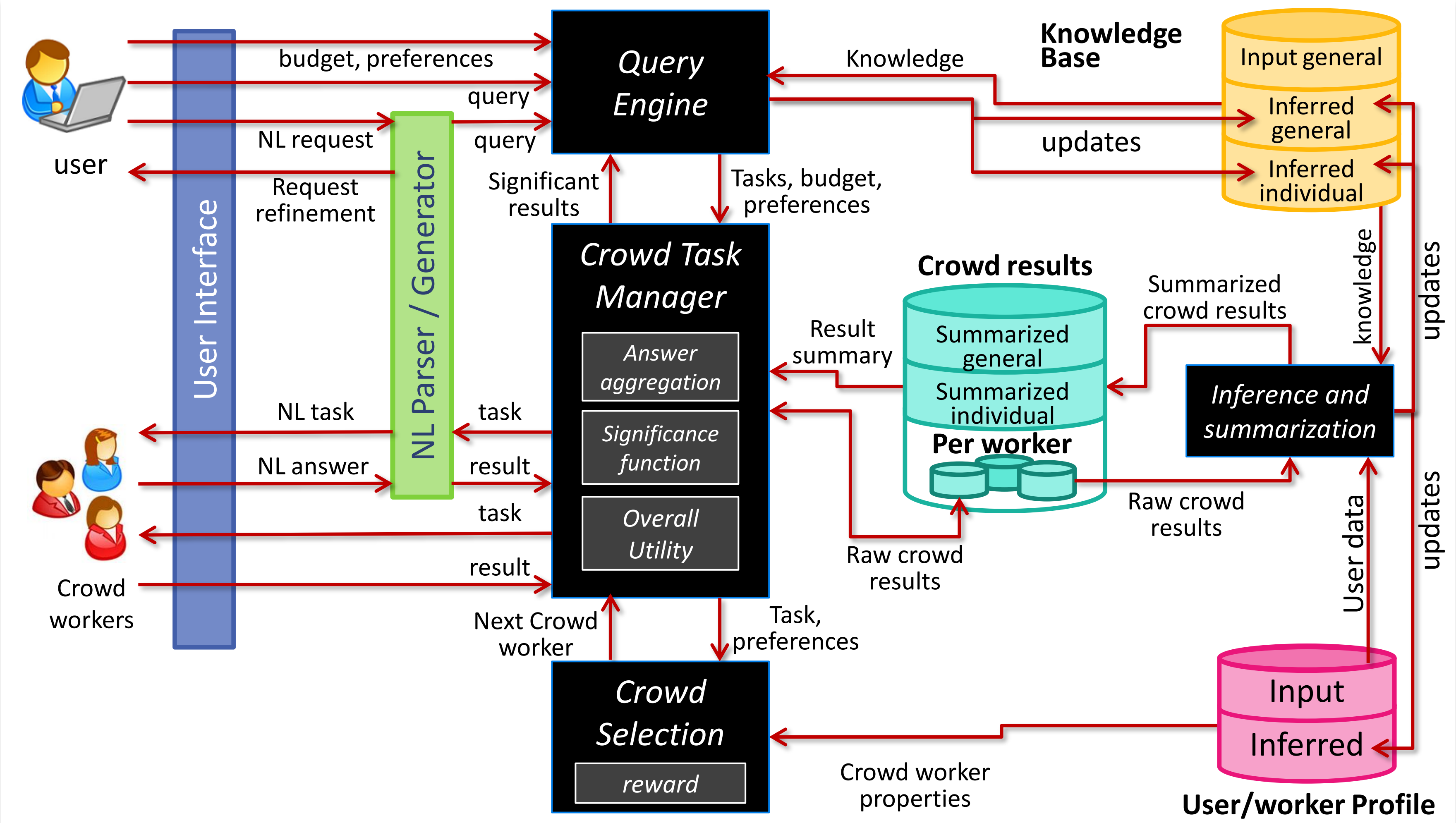
Motivation

Framework Architecture

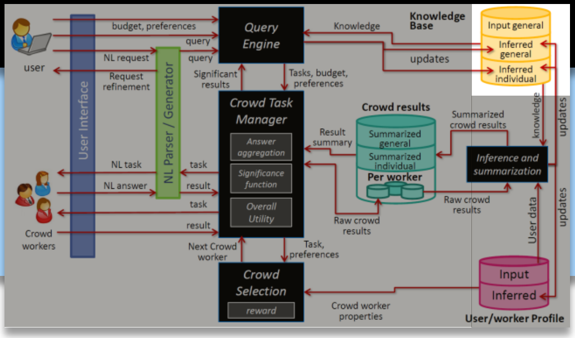
Zoom-in on components

Examples via the OASSIS system

The Architecture



Knowledge Repository

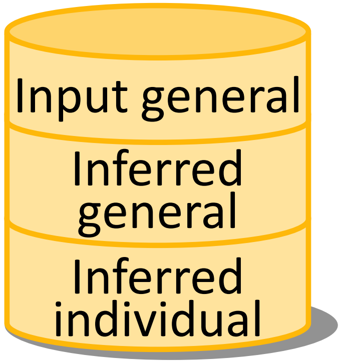


Different types of knowledge:

- A general knowledge base is input to the system

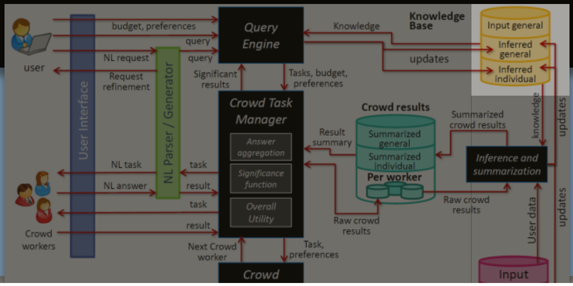
WordNet foursquare yago select knowledge

- Knowledge inferred in previous query evaluation
  - **General knowledge** – completes the knowledge base
    May be annotated with trust/error probability
  - **Individual knowledge** – more volatile
    may be annotated with user properties

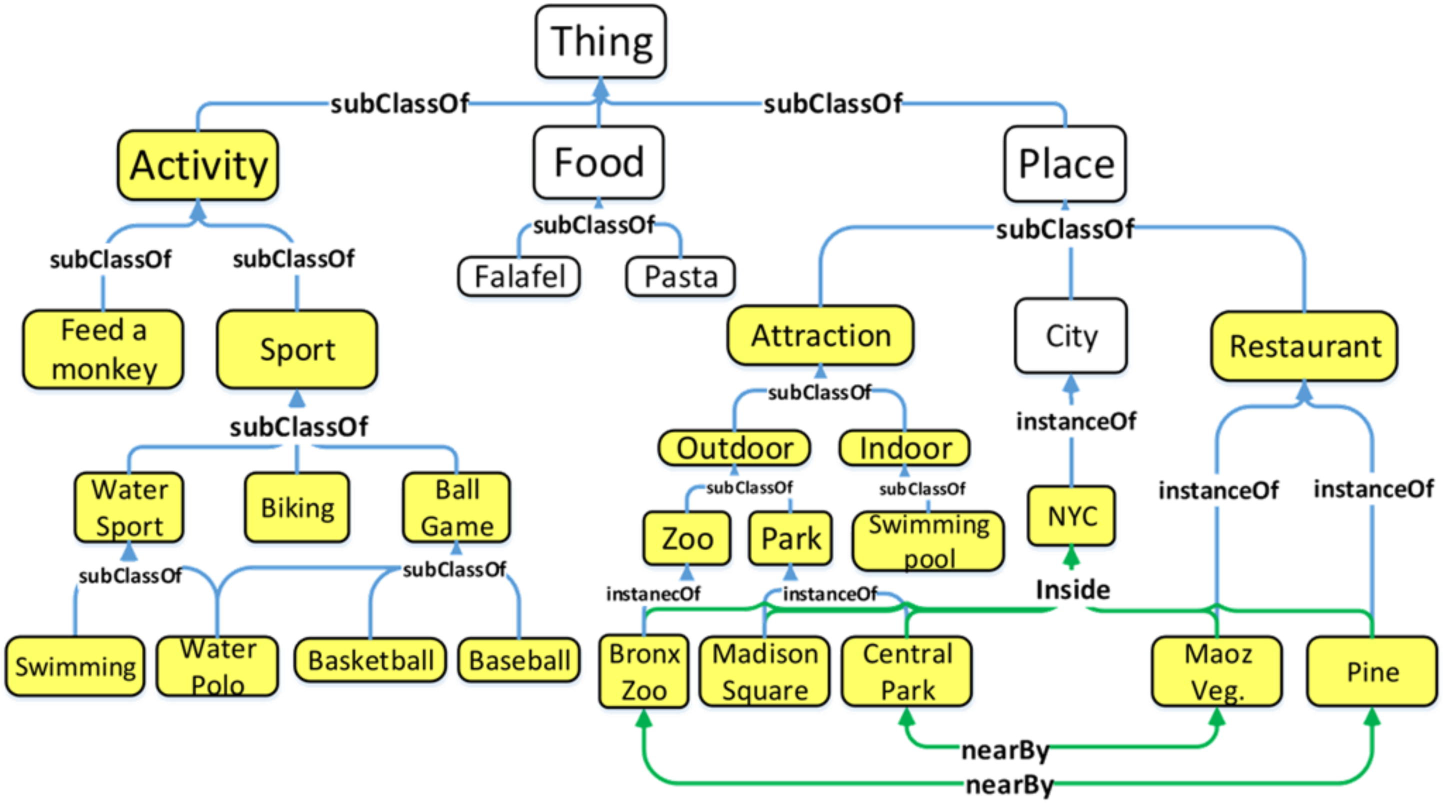

Input general

Inferred general

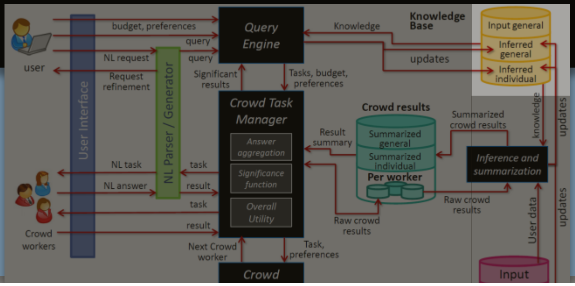
Inferred individual

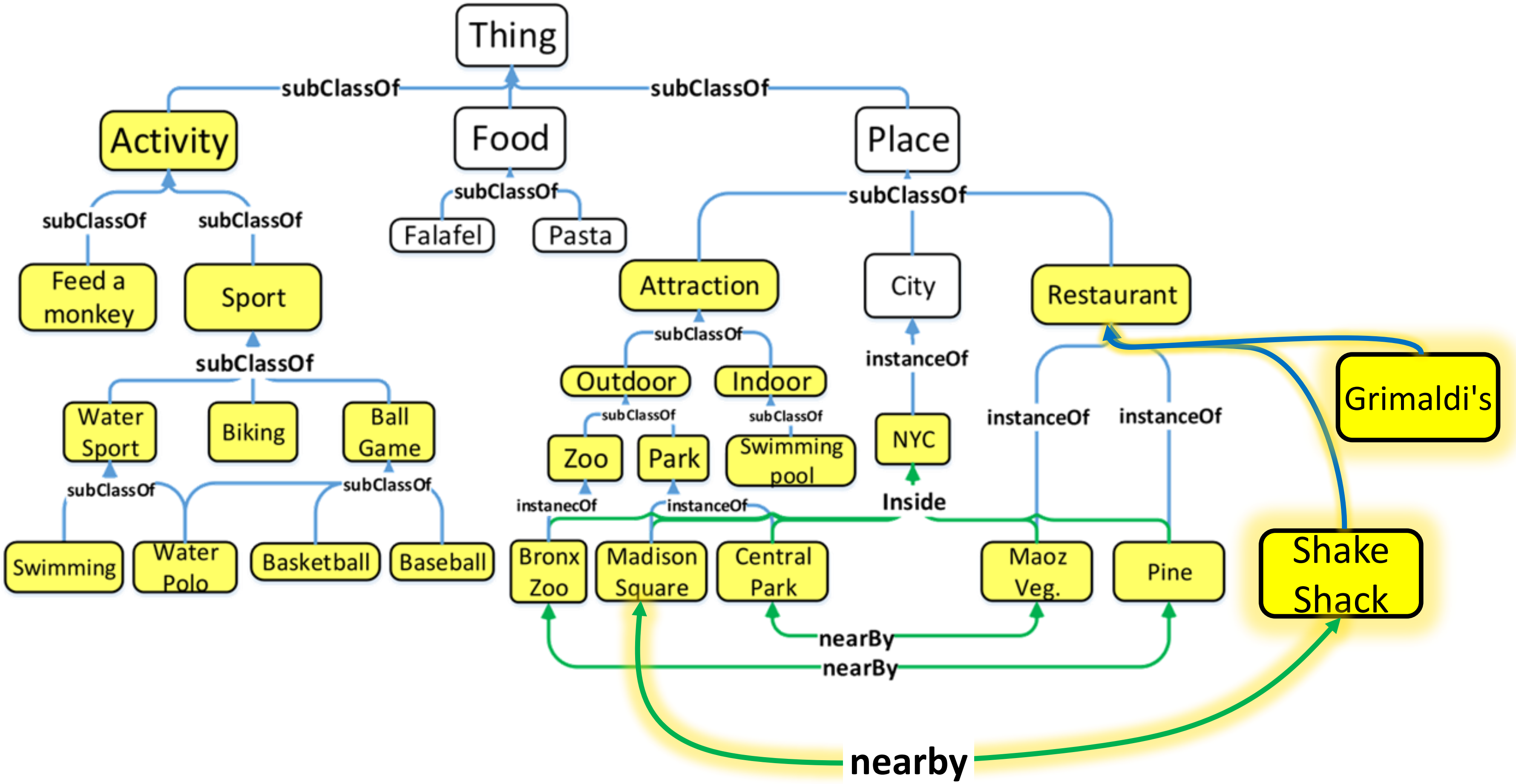


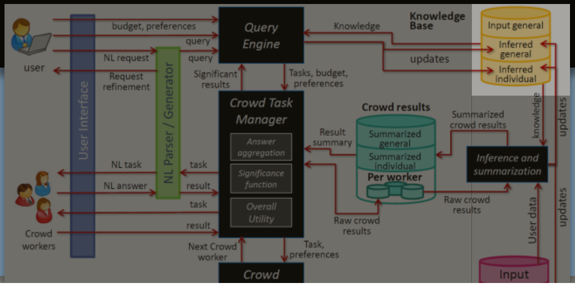
Knowledge Repository



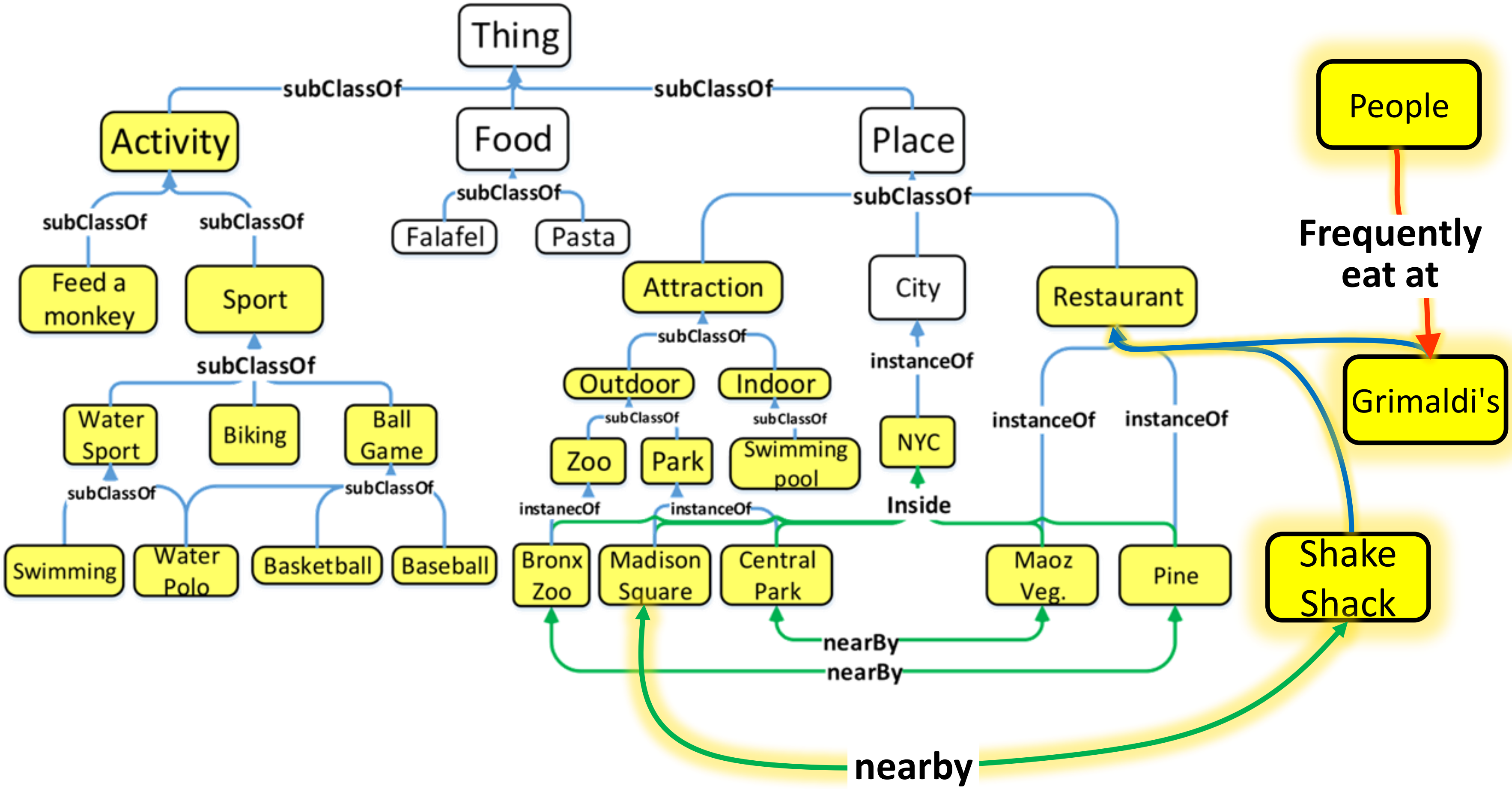
Knowledge Repository



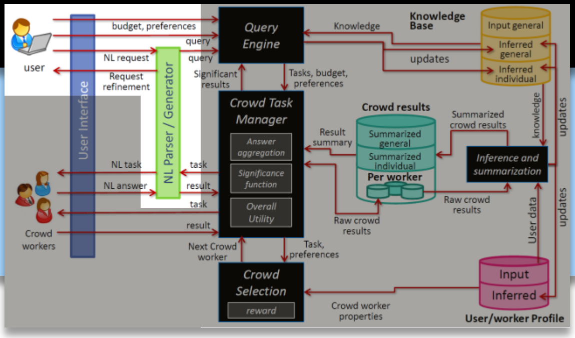




Knowledge Repository



Enters the user…



- The user query should be formulated in a formal language

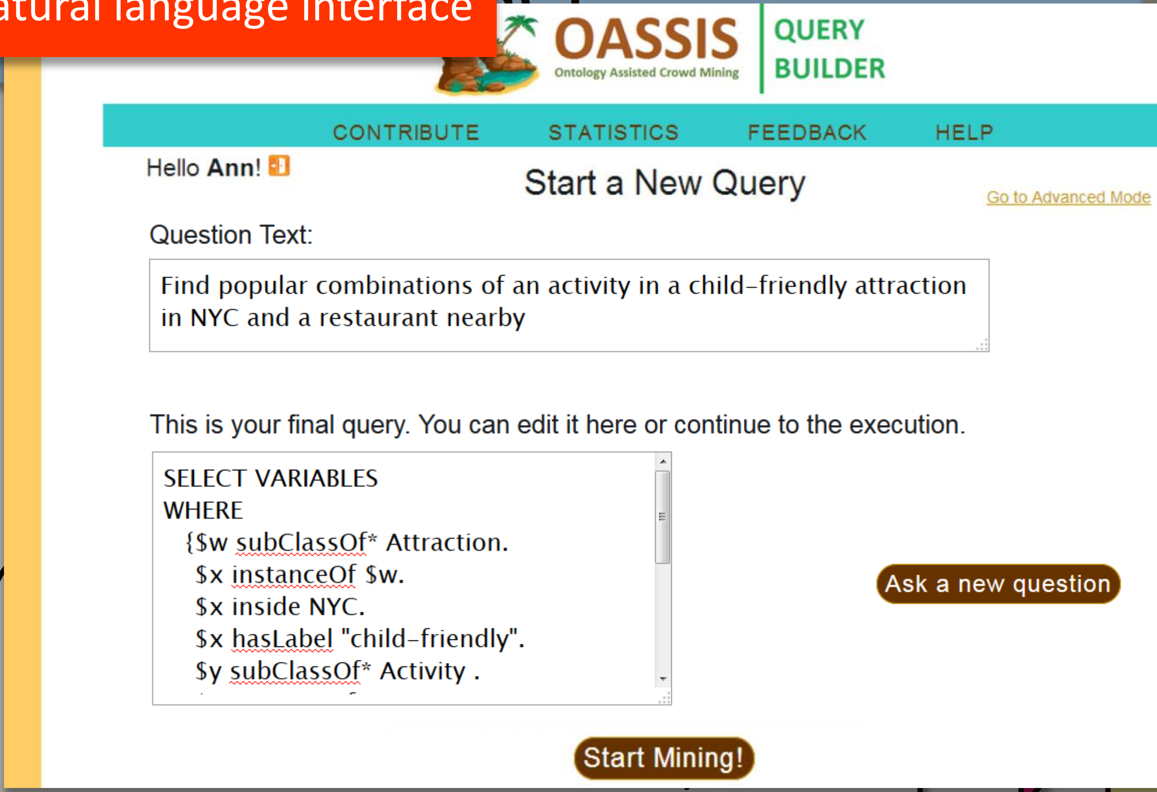
  E.g., OASSIS-QL is a SPARQL-based query language for crowd mining

  [A. et al. SIGMOD’14]

Find popular combinations of
an activity in a child-friendly attraction
at NYC and a restaurant nearby
(plus relevant advice)

```
SELECT VARIABLES
WHERE
     {$w subClassOf* Attraction
      $x instanceOf $w.
      $x inside      NYC.
      $y subClassOf* Activity.
      $z instanceOf Restaurant.
      $z nearBy      $x}
SATISFYING
      {$y+ doAt      $x.
      []   eatAt     $z.
      MORE}
      WITH SUPPORT = 0.03
```

Natural language interface



Hello **Ann!**

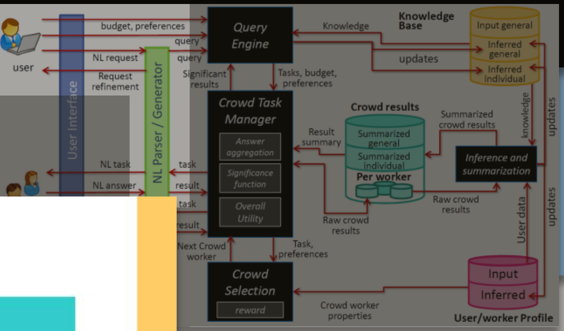
Start a New Query

Question Text:

Find popular combinations of an activity in a child-friendly attraction in NYC and a restaurant nearby

This is your final query. You can edit it here or continue to the execution.

```
SELECT VARIABLES
WHERE
  {$w subClassOf* Attraction.
   $x instanceOf $w.
   $x inside NYC.
   $x hasLabel "child-friendly".
   $y subClassOf* Activity .
```



ser

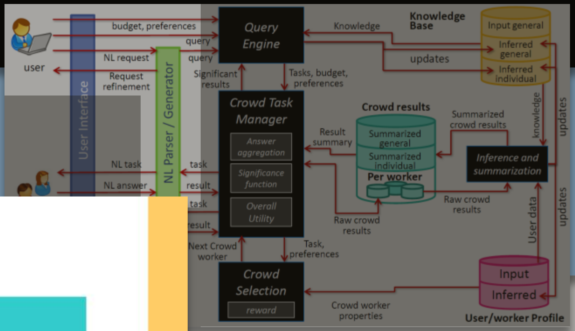
l language

wd mining

(plus relevant advice)

```
   IABLES

   bClassOf* Attraction
   stanceOf $w.
   side      NYC.
   bClassOf* Activity.
   stanceOf Restaurant.
 $z nearBy    $x}
SATISFYING
  {$y+ doAt      $x.
   []   eatAt    $z.
   MORE}
   WITH SUPPORT = 0.03
```

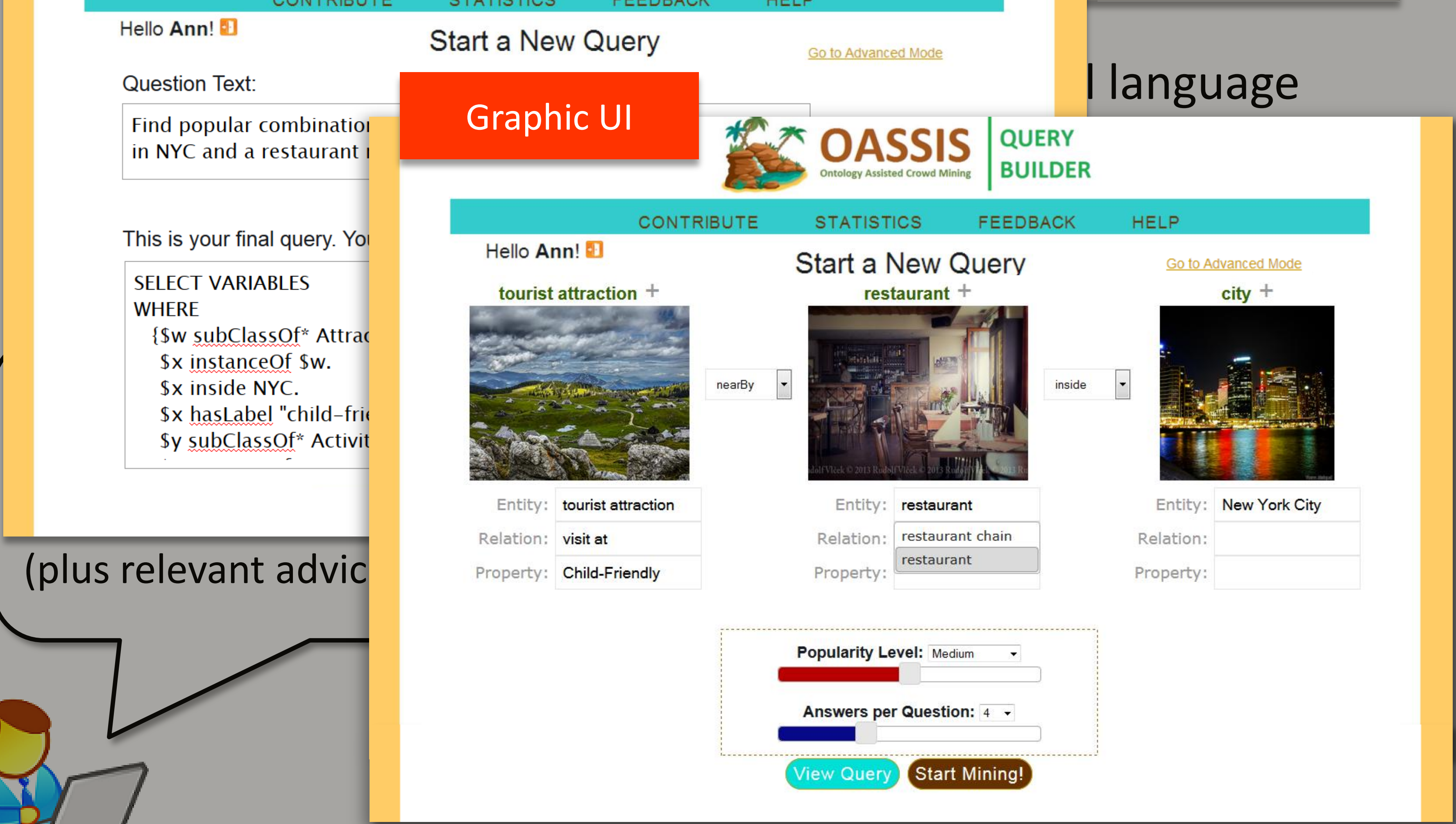


...ser

...l language

Natural language interface

Graphic UI



Hello Ann!

Start a New Query

Question Text:

Find popular combinatio...
in NYC and a restaurant ...

This is your final query. Yo...

```
SELECT VARIABLES
WHERE
  {$w subClassOf* Attrac...
   $x instanceOf $w.
   $x inside NYC.
   $x hasLabel "child-fri...
   $y subClassOf* Activit...
```

Hello Ann!

Start a New Query

Go to Advanced Mode

tourist attraction +		restaurant +		city +
	nearBy		inside	

Entity:	tourist attraction	Entity:	restaurant	Entity:	New York City
Relation:	visit at	Relation:	restaurant chain restaurant	Relation:	
Property:	Child-Friendly	Property:		Property:	

Popularity Level: Medium

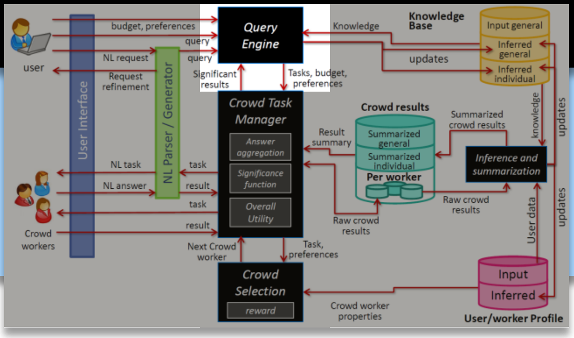
Answers per Question: 4

View Query

Start Mining!

(plus relevant advic...

Query Engine

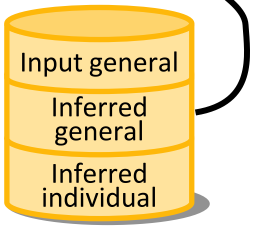


- Efficiently executes the query plan
  - By querying the knowledge base (standard)
  - And generating questions/tasks to the crowd

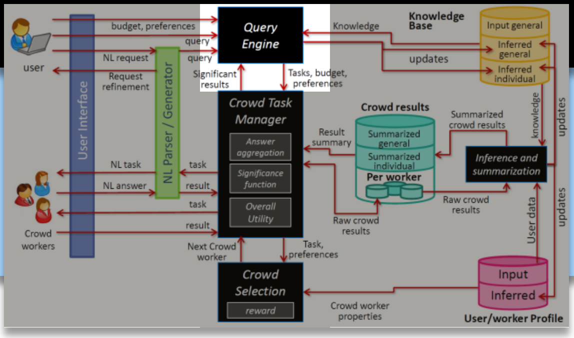
```
{$x instanceOf    Attraction.
 $y subClassOf    Activity}
```

```
{$y  doAt          $x}
```

```
$x = Central_Park
$y = Baseball
```



Crowd task:
isSignificant({Baseball doAt Central_Park})
Budget: $0.5
User preferences: …



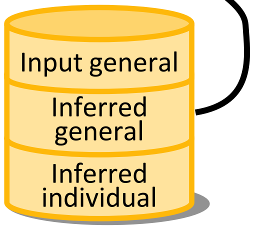
Query Engine

- Efficiently executes the query plan
  - By querying the knowledge base (standard)
  - And generating questions/tasks to the crowd

```
{$x instanceOf    Attraction.
 $y subClassOf    Activity}
```

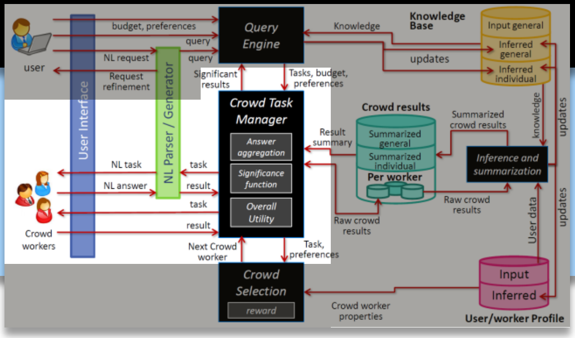
```
{$y  doAt          $x}
```

```
$x = Central_Park
$y = Baseball
```



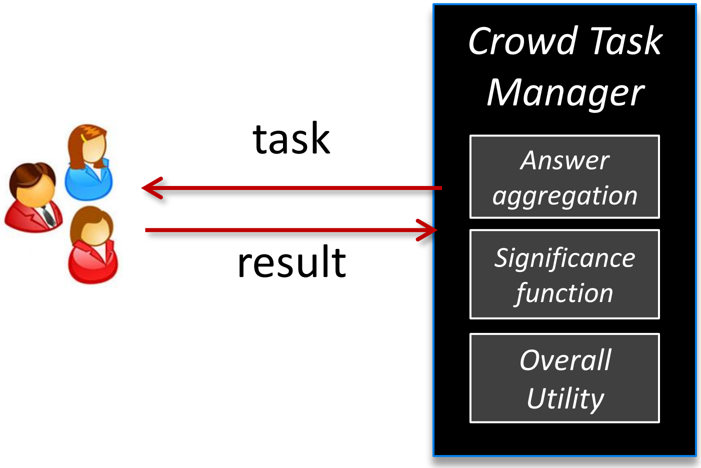
Crowd task:
isSignificant({Baseball doAt Central_Park})
Bu
U

Crowd task:
specify($z, {Baseball doAt Central_Park.
            [] eatAt $z})
Budget: $0.6



Crowd Task Manager

- Distributes tasks to crowd members
- Aggregates and analyzes the answers
- Dynamically decides what to ask next


task

result

Crowd Task Manager

Answer aggregation

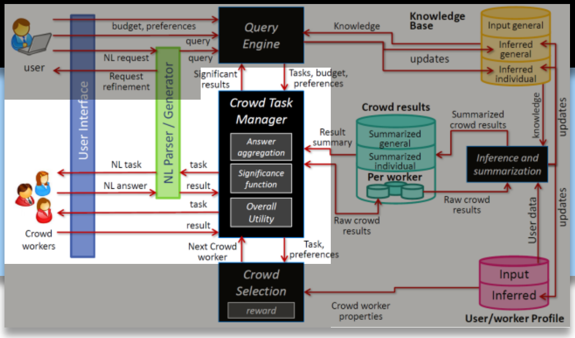
Significance function

Overall Utility

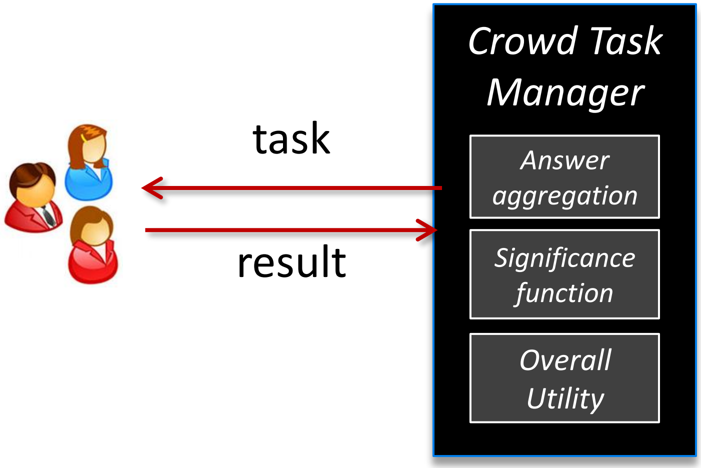
Crowd task:
isSignificant(`{Baseball doAt Central_Park}`)
Budget: $0.5
User preferences: …

“How often do you play baseball at Central Park?”

Crowd Task Manager



- Distributes tasks to crowd members
- Aggregates and analyzes the answers
- Dynamically decides what to ask next



task

result

Crowd Task Manager

Answer aggregation

Significance function

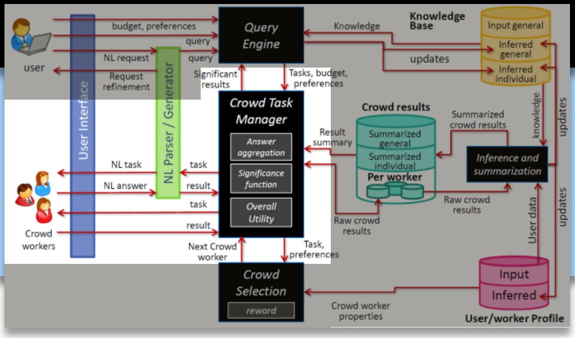
Overall Utility

Crowd task:
isSignificant(`{Baseball doAt Central_Park}`)
Budget: $0.5
User preferences: …

"How often do you play baseball at Central Park?"

Answer 1: never (score=0)

Crowd Task Manager



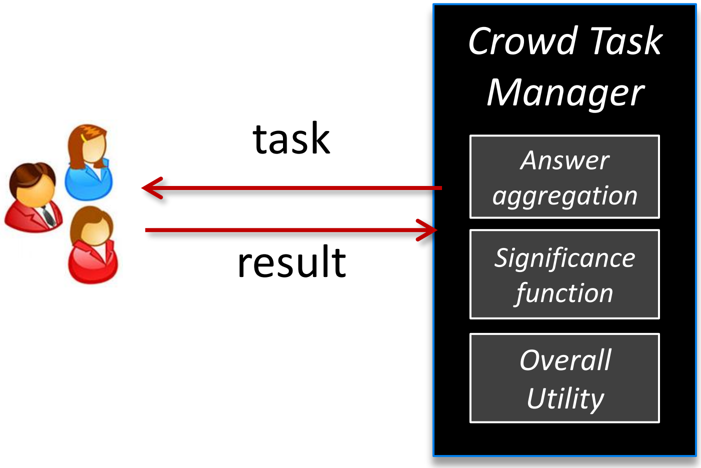
- Distributes tasks to crowd members
- Aggregates and analyzes the answers
- Dynamically decides what to ask next

Crowd task:
isSignificant(`{Baseball doAt Central_Park}`)
Budget: $0.5
User preferences: …

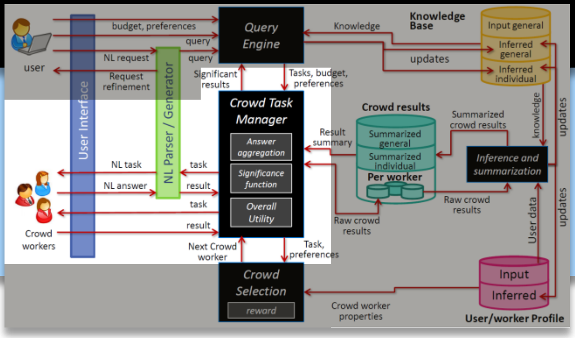
“How often do you play baseball at Central Park?”

Answer 1: never (score=0)

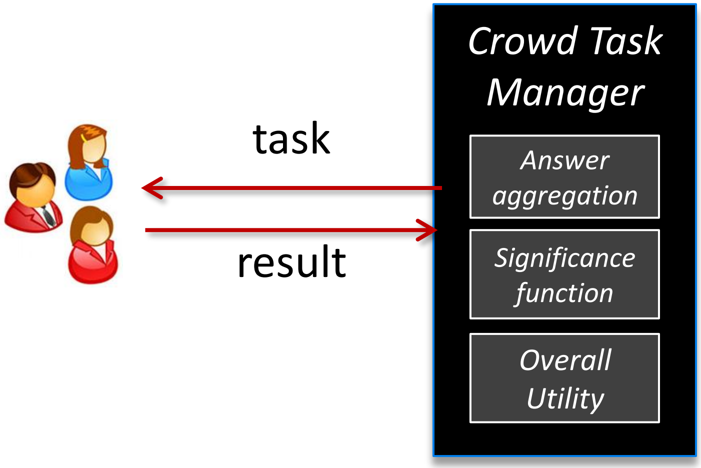
Answer 2: once a week (score=1/7)



Crowd Task Manager



- Distributes tasks to crowd members
- Aggregates and analyzes the answers
- Dynamically decides what to ask next



task

result

Crowd Task Manager

Answer aggregation

Significance function

Overall Utility

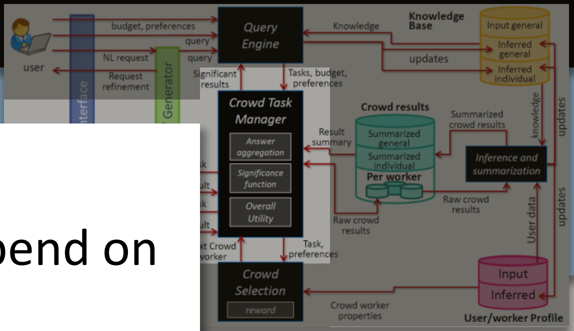
Crowd task:
isSignificant({Baseball doAt Central_Park})
Budget: $0.5
User preferences: …

"How often do you play baseball at Central Park?"

Answer 1: never (score=0)

Answer 2: once a week (score=1/7)

Aggregation: estimated mean M
Significance: $\Pr(M \geq \Theta) \geq 0.5$
Overall utility: next question expected to reduce error probability by 0.1



C

Aggregation, significance and utility choices depend on the type of data collected from the crowd.

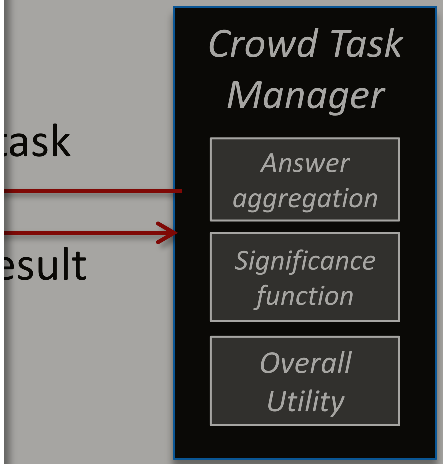
For **individual** data, the aggregated answer should account for diverse opinions

- e.g., statistical modeling

For **general** data the aggregated answer should reflect the truth

- e.g., weighing by expertise, outlier filtering

C
is
B
U


ask

esult

Crowd Task Manager

Answer aggregation

Significance function

Overall Utility

“How often do you play baseball at Central Park?”

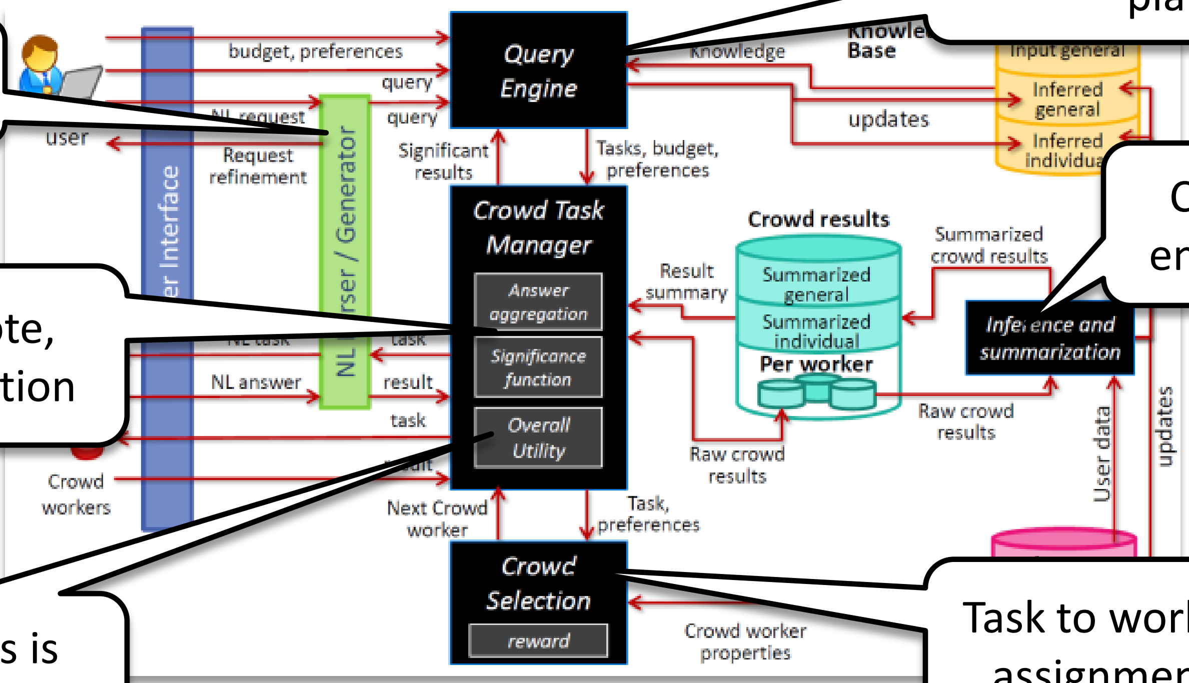
Answer 1: never (score=0)

Answer 2: once a week (score=1/7)

Aggregation: estimated mean M
Significance: $\Pr(M \geq \Theta) \geq 0.5$
Overall utility: next question expected to reduce error probability by 0.1

Other crowdsourcing systems

Can be put in terms of the architecture for comparing and identifying possible extensions



NL to query translators

Majority vote, custom function

# questions is fixed or bounded

Declarative crowdsourcing platforms

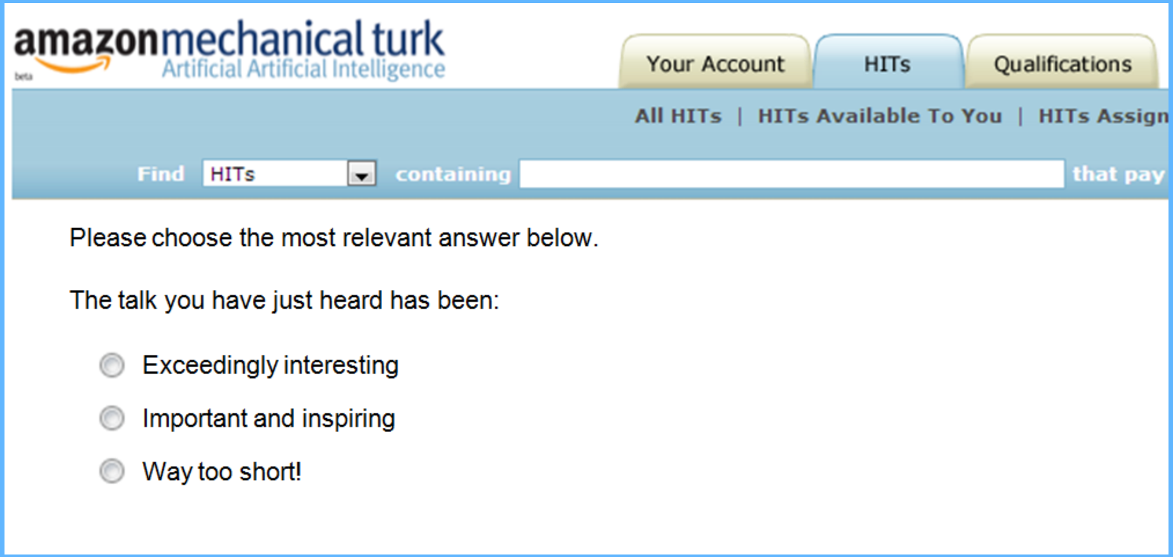
Crowdsourced entity resolution

Task to worker assignment

In Conclusion

- Crowd mining allows users to ask queries that mix general and individual data needs,
  and use multiple sources to obtain relevant answers
- Our generic architecture outlines the components required for such complex reasoning
- Other crowdsourcing systems share a part of these components, possibly with alternative implementations
- This analysis highlights challenges for future work

Thank you



amazonmechanical turk
Artificial Artificial Intelligence
beta
Your Account
HITs
Qualifications
All HITs | HITs Available To You | HITs Assign
Find HITs containing that pay

Please choose the most relevant answer below.

The talk you have just heard has been:

○ Exceedingly interesting

○ Important and inspiring

○ Way too short!