# Dr. Detective

combining gamification techniques and crowdsourcing to create a gold standard in medical text

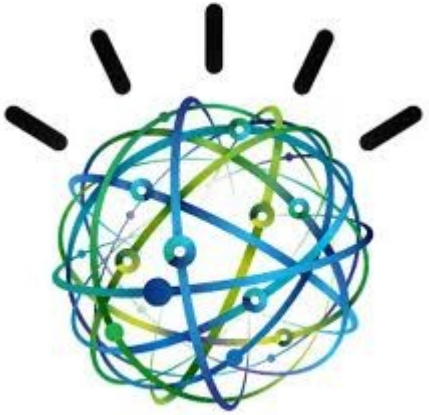
21/10/2013

Anca Dumitrache, Lora Aroyo, Chris Welty, Robert-Jan Sips, Antony Levas

# Ground truth for Watson

- **Watson =** natural language question answering system
- *training data:*
  - databases
  - ontologies
  - taxonomies

  i.e. **ground truth**



- *goal:* adaptation of Watson for other domains

  (e.g. **medical domain**)
- *problem:* how to acquire ground truth?

# Ground truth for Watson

Training Watson for the medical domain:

- Answer questions about diagnosing
- Find synonym phrases
- Identify negation (and its variations)
- Identify different term types
- Identify relations between term types

# Ground truth for Watson

Training Watson for the medical domain:

- **Answer questions about diagnosing**
- Find synonym phrases
- Identify negation (and its variations)
- **Identify different term types**
- Identify relations between term types

# Ground truth for Watson

*Problem:* language ambiguity

# Ground truth for Watson

*Problem:* language ambiguity

Patients exhibiting *acute tailbone pain* should be examined for extra bone nodules.

What is the primary term?

# Ground truth for Watson

*Problem:* language ambiguity

Patients exhibiting *acute tailbone* ***pain*** should be examined for extra bone nodules.

What is the primary term?

# Ground truth for Watson

*Problem:* language ambiguity

Patients exhibiting *acute* ***tailbone pain*** should be examined for extra bone nodules.

What is the primary term?

# Ground truth for Watson

*Problem:* language ambiguity

Patients exhibiting ***acute tailbone pain*** should be examined for extra bone nodules.

What is the primary term?

# Ground truth for Watson

*Problem:* language ambiguity

Patients exhibiting ***acute tailbone pain*** should be examined for extra bone nodules.

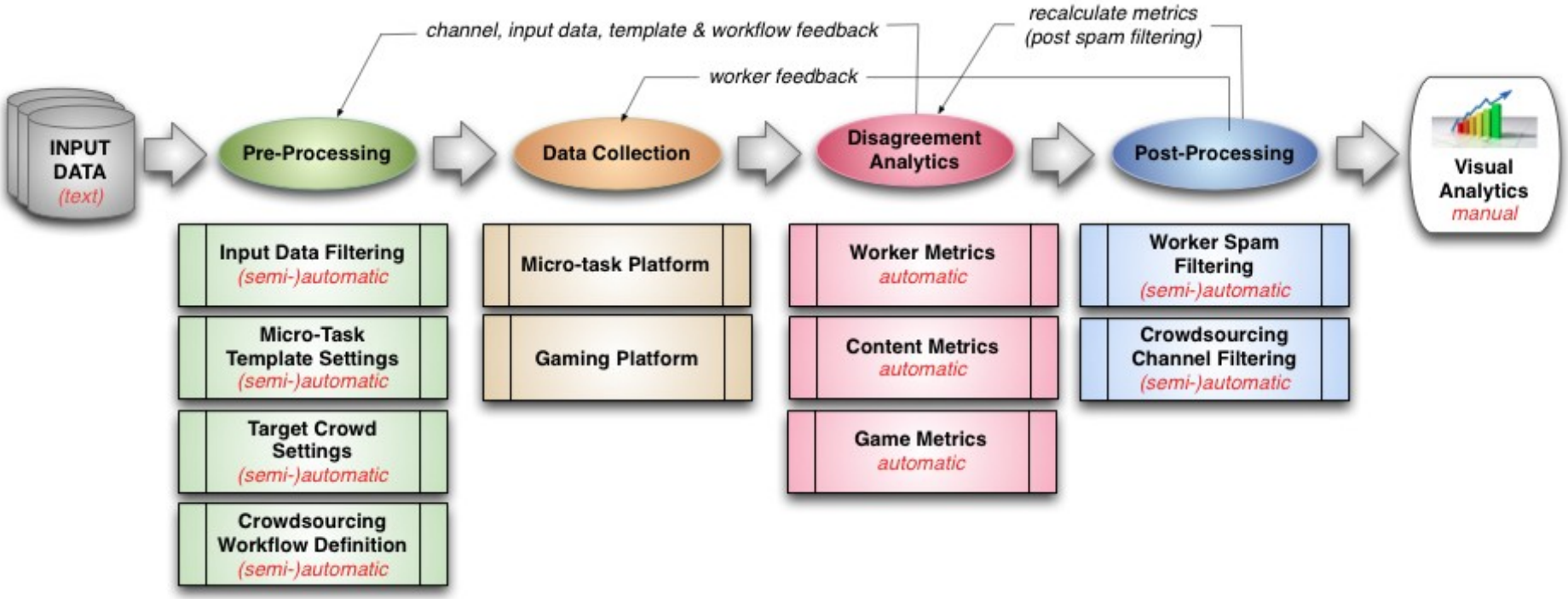
What is the primary term?

- *traditional approach:* guidelines for consistently choosing one answer (e.g. ***tailbone pain*** is the primary term)
- ***Crowd Truth*** *approach:* capture and measure diversity of opinion (e.g. by counting votes for each variation of a term)

# Research goals

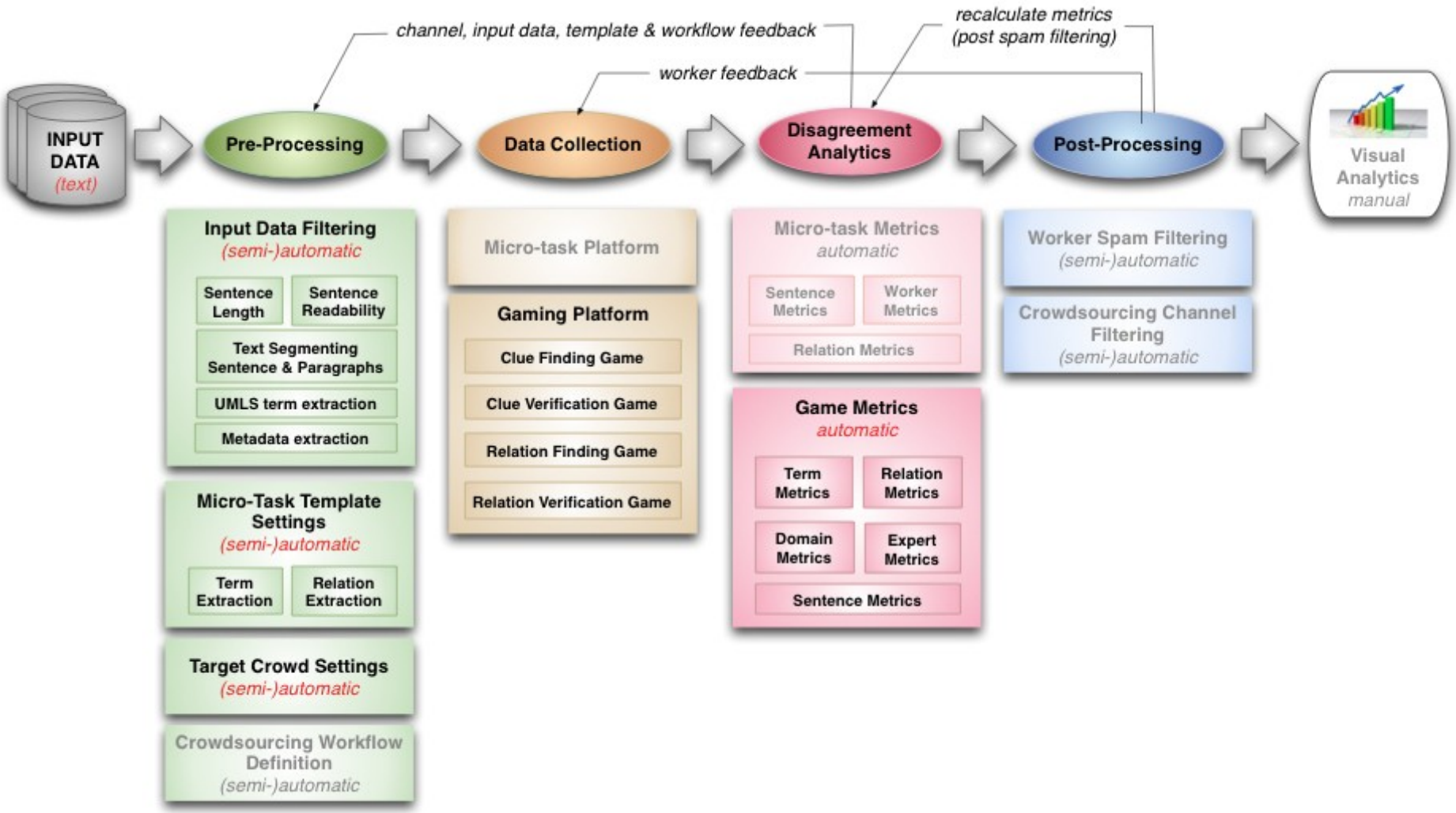
- Investigate the feasibility of a game for niche (expert) sourcing that captures a diversity of opinions
- Measure quality of *Crowd Truth* through metrics
- Evaluate effectiveness of game for engaging the experts

# *Crowd-Watson* Workflow




available at: http://crowd-watson.nl

# *Crowd-Watson* Workflow: Game Aspects




available at: http://crowd-watson.nl/dr-detective-game/

# *Crowd-Watson* Workflow: Input Data

- *input:* patient case reports
- *source:* New England Journal of Medicine

CASE RECORDS OF THE MASSACHUSETTS GENERAL HOSPITAL

Richard C. Cabot, Founder, Eric S. Rosenberg, M.D., Editor, Nancy Lee Harris, M.D., Editor, Jo-Anne O. Shepard, M.D., Associate Editor, Alice M. Cort, M.D., Associate Editor, Sally H. Ebeling, Assistant Editor, Emily K. McDonald, Assistant Editor

## Case 23-2013 — A 54-Year-Old Woman with Abdominal Pain, Vomiting, and Confusion

Kamyar Kalantar-Zadeh, M.D., M.P.H., Ph.D., Raul N. Uppot, M.D., and Kent B. Lewandrowski, M.D.

N Engl J Med 2013; 369:374-382 | July 25, 2013 | DOI: 10.1056/NEJMcpc1208154

Comments open through July 31, 2013

Poll open through July 23, 2013

Share:

Article | References | Comments (221)

*The description of this case was presented as a Case Challenge. Readers were invited to review the case description, vote on the diagnosis, and submit comments. The full case discussion and final diagnosis now appear below, along with the poll results.*

### PRESENTATION OF CASE

*Dr. Sara R. Schoenfeld* (Medicine): A 54-year-old woman was admitted to this hospital because of abdominal pain, vomiting, and confusion.

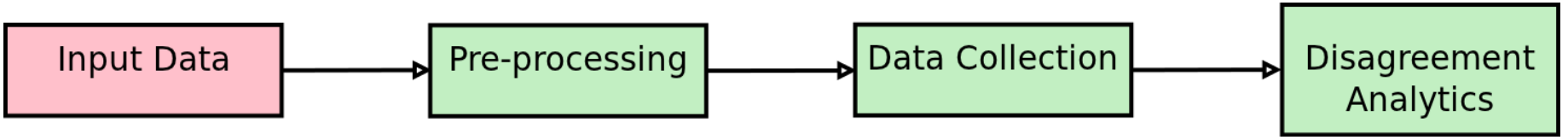
The patient was in her usual health until approximately 3 days before admission, when she reportedly began to feel unwell, with weakness, chills, and skin that was abnormally warm to the touch. She self-administered aspirin, without improvement. During the next 2 days, her oral intake decreased. Approximately 22 hours before presentation, vomiting occurred. Nine hours before presentation, she began to travel home to Italy from the eastern United States. During the next 2 hours, increasing abdominal pain occurred, associated with vomiting and shortness of breath, and

. . .

notification from her primary care doctor 1 week later that she was doing well and had resumed her normal daily activities.

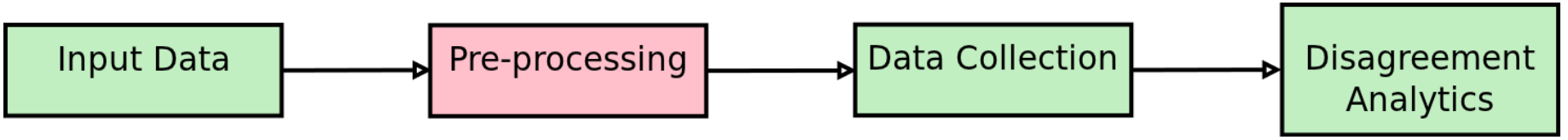
### FINAL DIAGNOSIS

Toxic effects of metformin.



# *Crowd-Watson* Workflow: Pre-processing

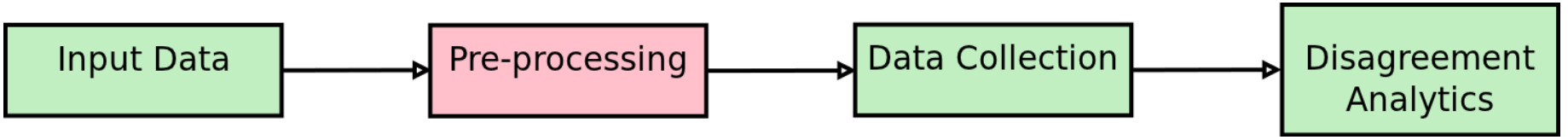
- input data filtering
- micro-task template setting
- target crowd setting



# *Crowd-Watson* Workflow: Pre-processing

- **input data filtering**
- micro-task template setting
- target crowd setting

- segment text into paragraphs
- extract diagnosis, specialization domains
- evaluate length of paragraphs, sentences
- evaluate readability (SMOG)
- medical named entity recognition (NER)



# *Dr. Detective* clue finding game

Dr. Watson | Home | Game | High Scores | About | Hi android! You scored: 63.0 points | (logout)

Level: **Hard**, Domain: **Primary Care/Hospitalist/Clinical Practice**

In the following text, find all the clues that could help diagnose **2009 influenza A (H1N1) virus infection**.

**Step 1: Select the type of clue you are looking for.**

Time/Duration

**Step 2: To pick a clue, highlight all the words that describe it by clicking on them.**

On the second day, hypoxemia (Table 2) and renal failure (Table 1) developed and urine output fell to 20 to 30 ml per hour. Transthoracic echocardiography showed an ejection fraction of 50% and was otherwise normal. Microscopical examination of the urine sediment revealed white-cell casts and granular casts, with tubular cells and nondysmorphic red cells. Continuous venovenous hemofiltration was begun, complicated by catheter-related thrombosis. Heparin was administered.

Show clues by others

**Step 3: After all the words in the clue are highlighted, save the clue.**

Save clue | per hour

On the second day ×

**Step 4: After you found all the clues for Time/Duration, submit them.**

Submit your clues for Time/Duration

**Step 5: Go back to step 1 and select another clue type, or move on to the next diagnosis.**

Next diagnosis >>

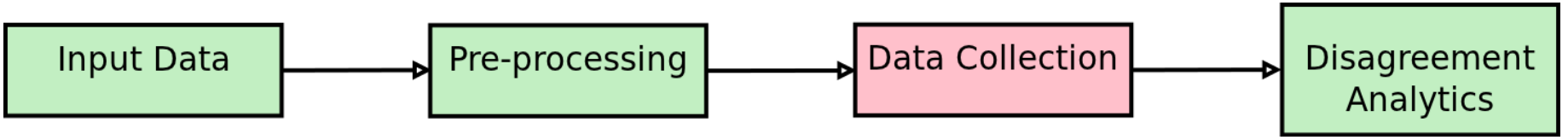
# *Dr. Detective* clue finding game

Gaming elements:

- **difficulty**
- scoring
- immersion
- playing options
- others' answers

Features in the difficulty vector of a paragraph:

- number of words in the paragraph
- number of sentences in the paragraph
- average sentence length
- SMOG readability index of text
- number of medical terms



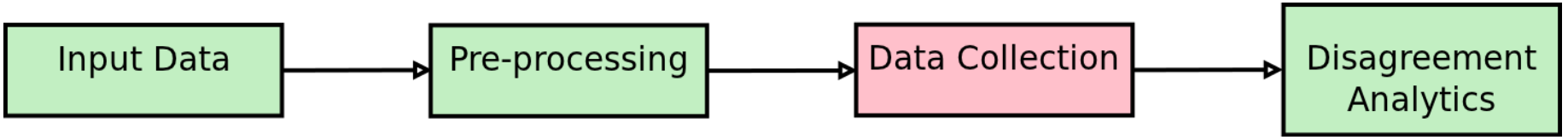
# *Dr. Detective* clue finding game

Gaming elements:

- difficulty
- **scoring**
- immersion
- playing options
- others' answers


Users receive points for:

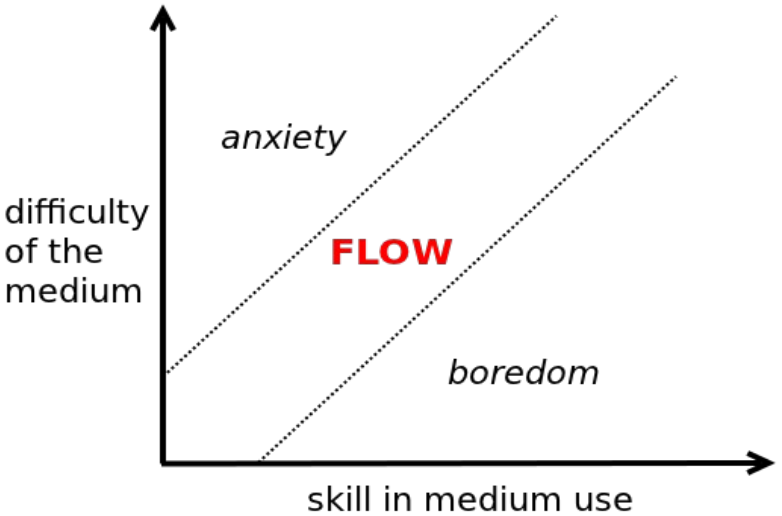
- paragraph difficulty
- consecutive answers
- popular answers
- new answers
- wrong answers



# *Dr. Detective* clue finding game

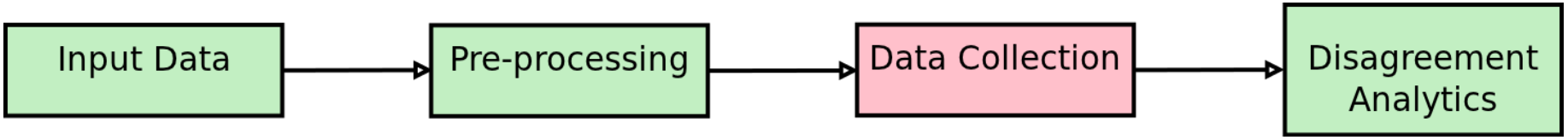
Gaming elements:

- difficulty
- scoring
- **immersion**
- playing options
- others' answers




*Sherry, "Flow and media enjoyment" (2004)*

- influences the next document selection mechanism
- next document needs to have increased difficulty, but minimum differing features



# *Dr. Detective* clue finding game

Gaming elements:

- difficulty
- scoring
- immersion
- **playing options**
- others' answers

Pick your domain:

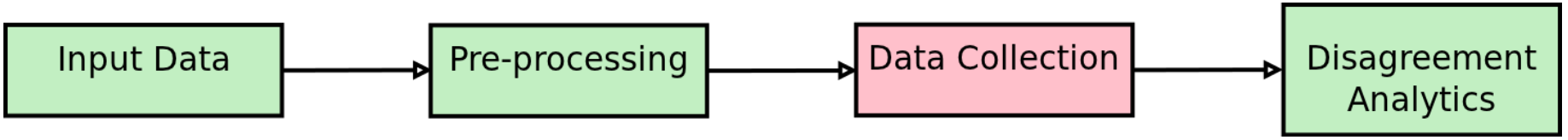
- Hematology/Oncology
- Nephrology
- Primary Care/Hospitalist/Clinical Practice
- Viral Infections

- **domains:** selected from the most popular categories in NEJoM

Pick the level you want to play:

| Quick Game | Normal Game | Hard Game |
|---|---|---|

- **levels:** based on the time it takes to solve a paragraph



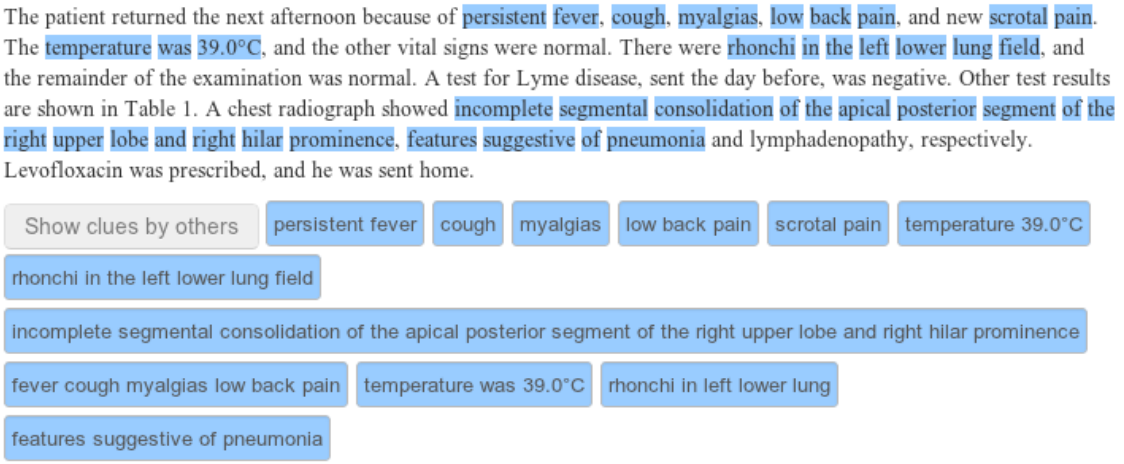
# *Dr. Detective* clue finding game

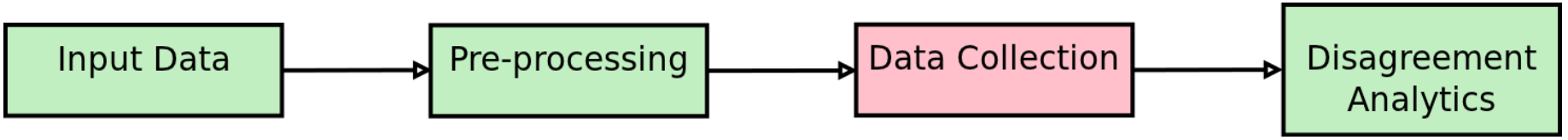
Gaming elements:

- difficulty
- scoring
- immersion
- playing options
- **others' answers**

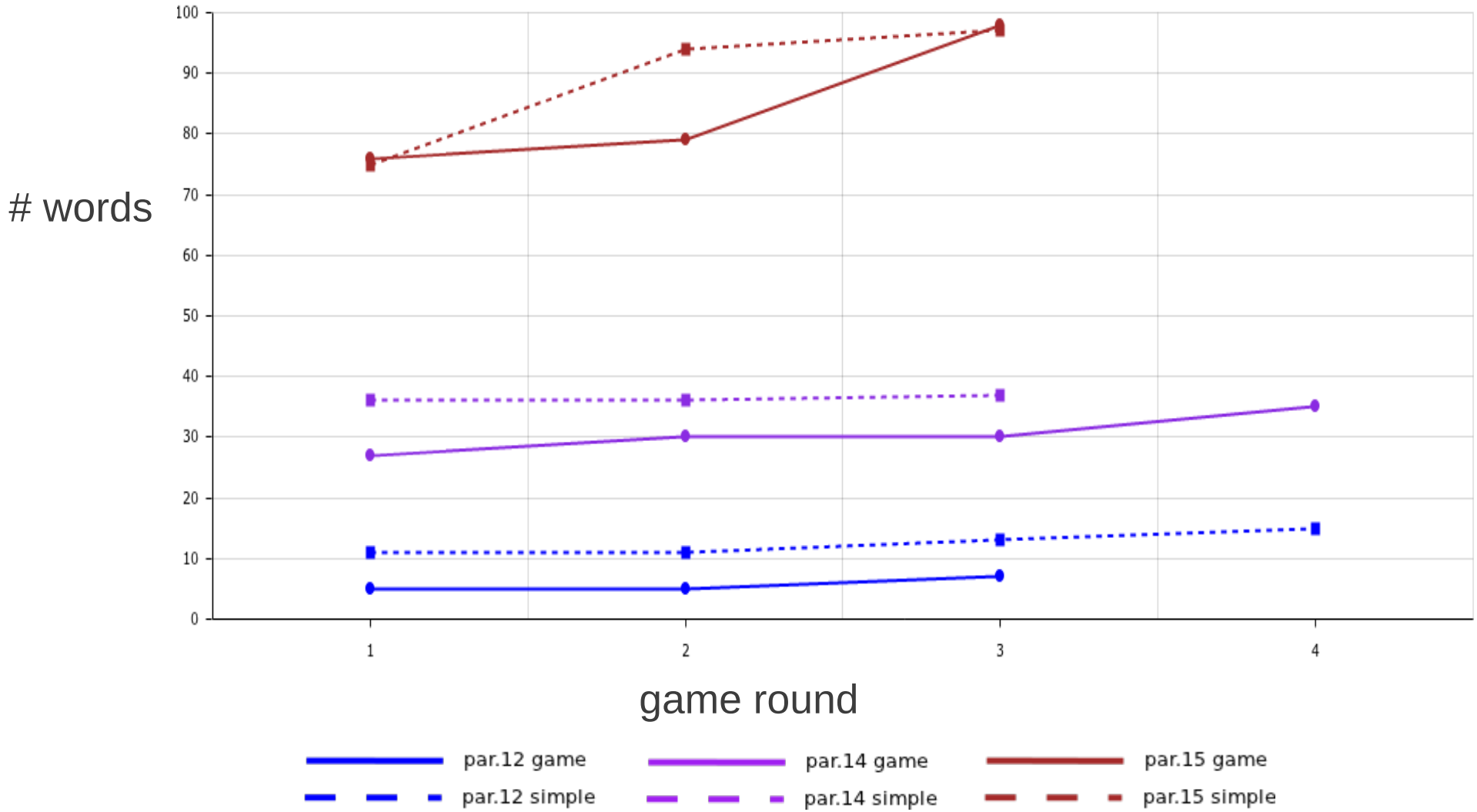
Q: *Does having access to the answers of other users stimulate diversity of opinion?*

- game v1: option to view others' answers
- game v2: no such option

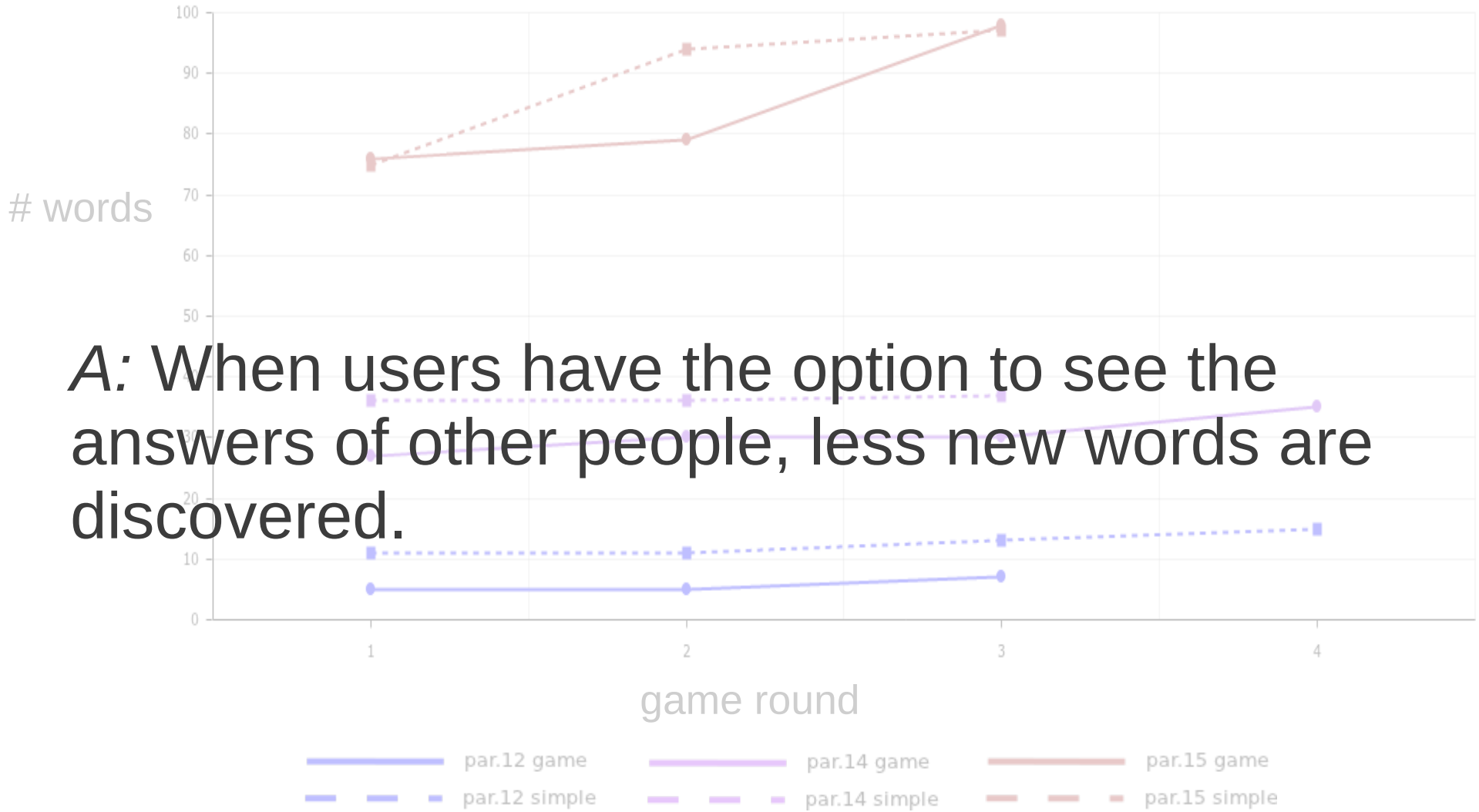


# *Q:* Does having access to the answers of other users stimulate diversity of opinion?



*Q:* Does having access to the answers of other users stimulate diversity of opinion?




*A:* When users have the option to see the answers of other people, less new words are discovered.

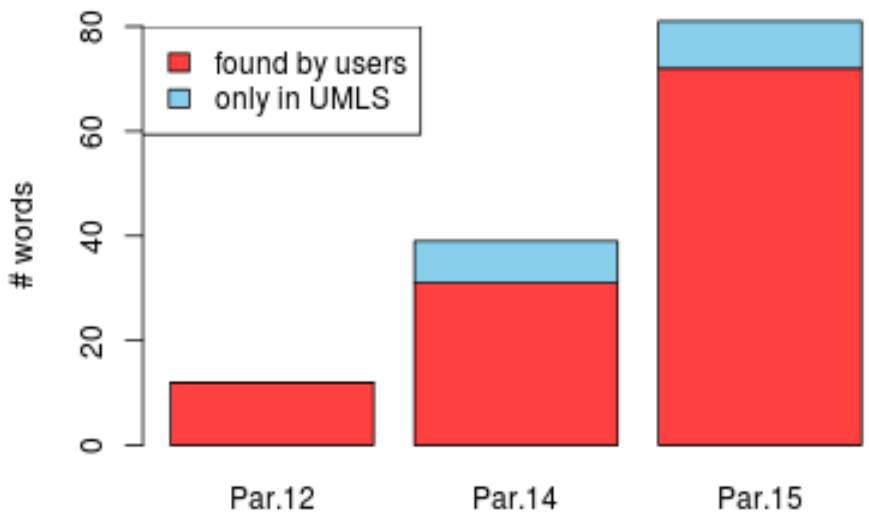
# Pilot run results

*Q:* How do the answers annotated by the crowd compare to those found by an NER?

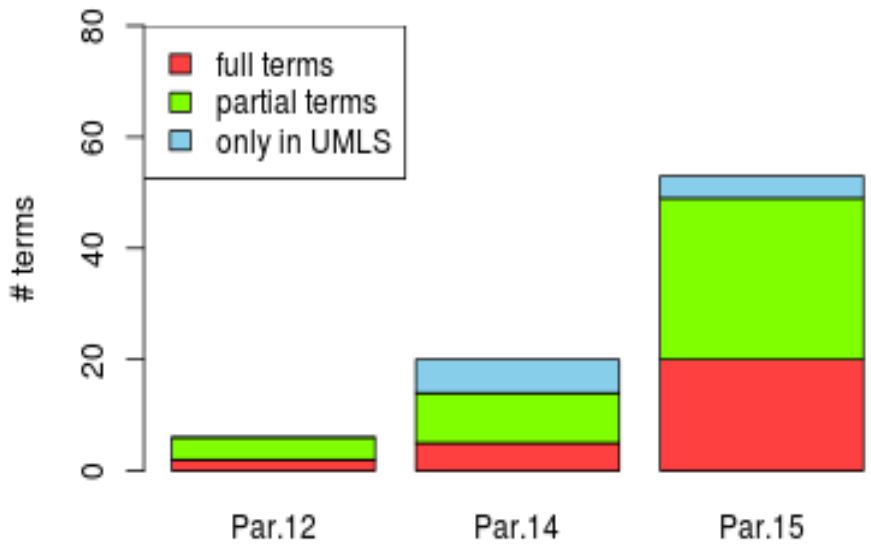
- NER: UMLS MetaMap
- full term match = the crowd found at least one term that has all the words in the term found by the NER
- partial term match = the crowd found at least one term that has any of the words in the term found by the NER

# Pilot run results

*Q:* How do the answers annotated by the crowd compare to those found by an NER?



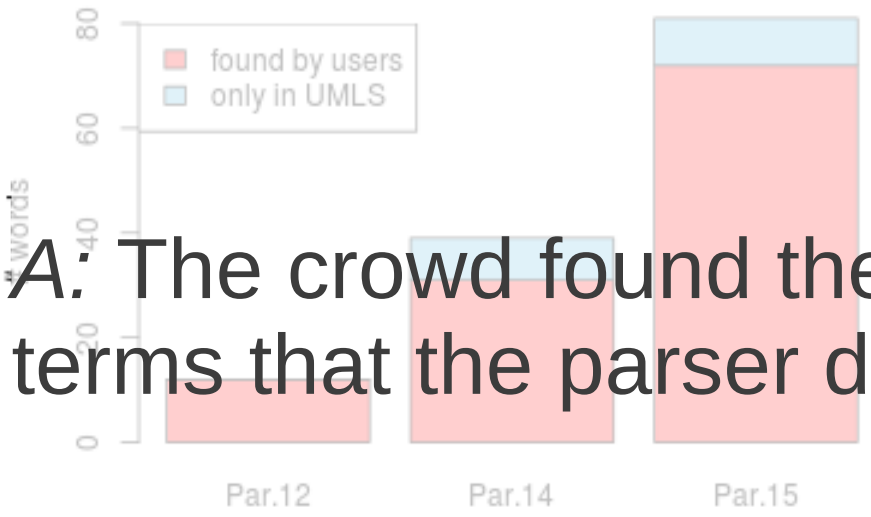

words distribution



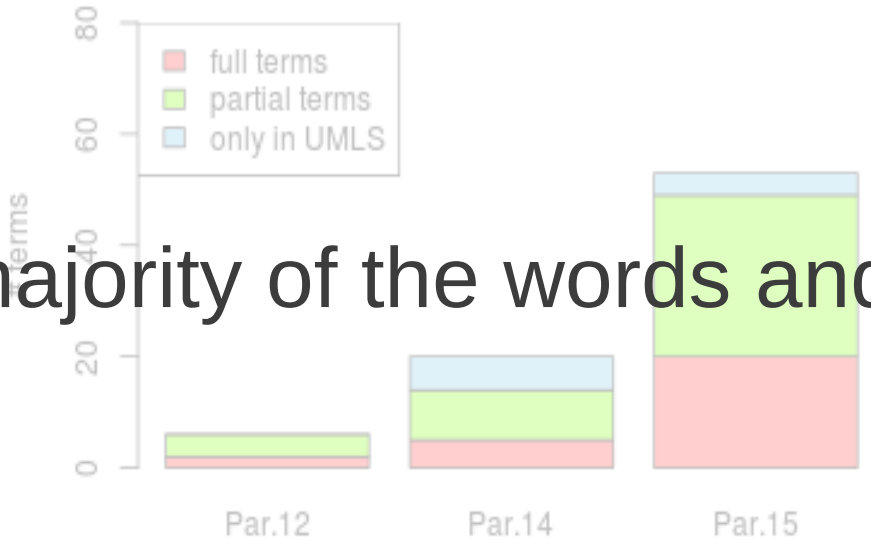

terms distribution

# Pilot run results

*Q:* How do the answers annotated by the crowd compare to those found by an NER?




words distribution



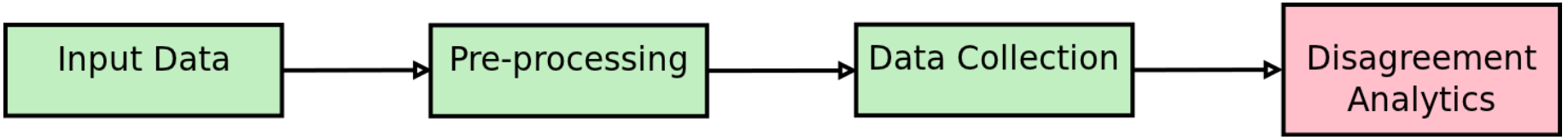

terms distribution

*A:* The crowd found the majority of the words and terms that the parser did.

# *Crowd-Watson* Workflow: Disagreement Analytics

Metrics:

- expert metrics
- sentence metrics
- term metrics
- relation metrics
- domain metrics



# Conclusions

- crowd answers are at least as accurate as an NER for term extraction and categorization
- to capture disagreement, access to the answers of others should be limited
- most users enjoyed playing the *Dr. Detective* game

# Future Work

- *user interaction:* score reports, more challenging tasks
- *data analysis:* specialized disagreement analytics for game
- *integration:* combine gaming and micro-task crowdsourcing
- running experiments with more participants

## Try it out:

- CrowdWatson: http://crowd-watson.nl
- Dr. Detective: http://crowd-watson.nl/dr-detective-game

## Related talks:

- *Content and Behavior-Based Metrics for Crowd Truth* @CrowdSem 14:45
- *Domain-Independent Quality Measure for Crowd Truth Disagreement* @DeRiVE 14:50

## Read more:

- *Measuring Crowd Truth for Medical Relation Extraction*, Aroyo & Welty, AAAI Fall Symposium SPD '13
- *Crowd Truth: Harnessing disagreement in crowdsourcing a relation extraction gold standard*, Aroyo & Welty, WebSci '13
- *Dr. Detective*, Dumitache, MSc thesis: http://goo.gl/doAERZ