# BioASQ

## A challenge on large-scale biomedical semantic indexing and question answering

George Paliouras and Anastasia Krithara

NCSR “D”

16th September 2014

BioASQ Workshop, Sheffield


Intelligent Information Management
Targeted Competition Framework
ICT-2011.4.4(d)

# Outline

# Introduction
## What is BioASQ

### A competition funded by the European Union (FP7)

- BioASQ initiates a series of **challenges** on **biomedical semantic indexing** and **question answering (QA)**.
- Participants are required to index semantically content form **large-scale** biomedical resources (e.g. MEDLINE) and/or
- to assemble data from **multiple heterogeneous sources** (e.g. scientific articles, knowledge bases, databases)
- to compose **informative answers** to biomedical natural language questions.

# Presentation of the challenge

## Tasks

### Task A: Hierarchical text classification

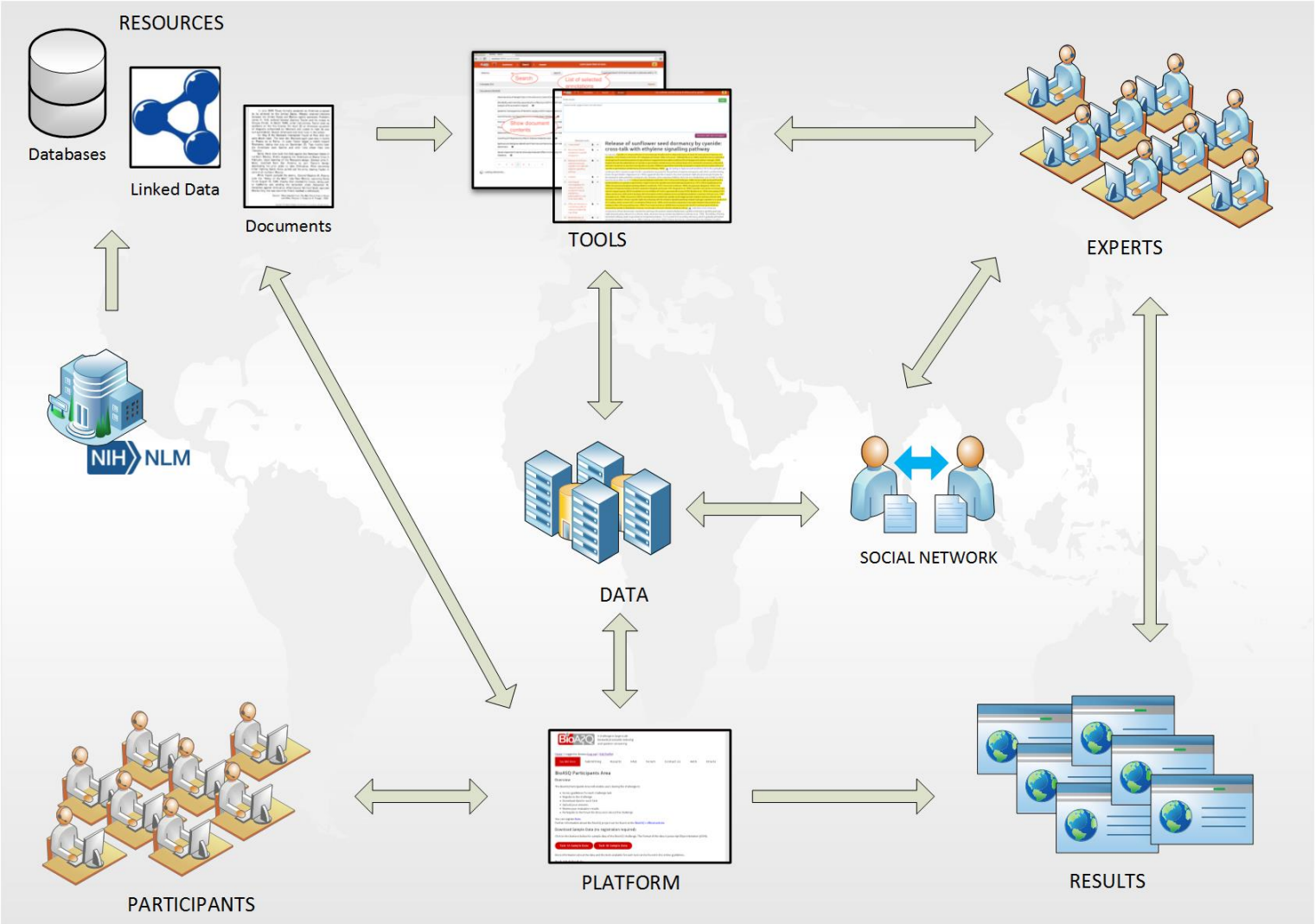
- Organizers distribute **new unclassfied PubMed articles.**
- Participants assign **MeSH terms** to the articles.
- **Evaluation** based on annotations of **PubMed curators.**

### Task B: IR, QA, summarization

- Organizers distribute **English biomedical questions.**
- Participants provide:relevant **articles, snippets, concepts, triples, exact answers, summary answers.**
- **Evaluation:** both **automatic** (GMAP, MRR, Rouge etc.) and **manual** (by biomedical experts).

# Presentation of the challenge

## Behind the scenes



# Presentation of the challenge

## Resources

### Criteria for selecting the resources

- **Publicly available**
- **Coverage** of different biomedical subfields
- Widely **acceptable** and **usable** format (e.g. OWL, OBO)
- **Low degree of overlap** between them

### Selected resources

- Data sources include both text and structured info:
  - Task 1a: Medline articles and MeSH
  - Task 1b:
    - PubMed abstracts and PubMed Central articles
    - Gene Ontology, UniProt, Jochem, Disease Ontology

# Presentation of the challenge

**What makes BioASQ more challenging:**

**LARGE SCALE data and knowledge sources**

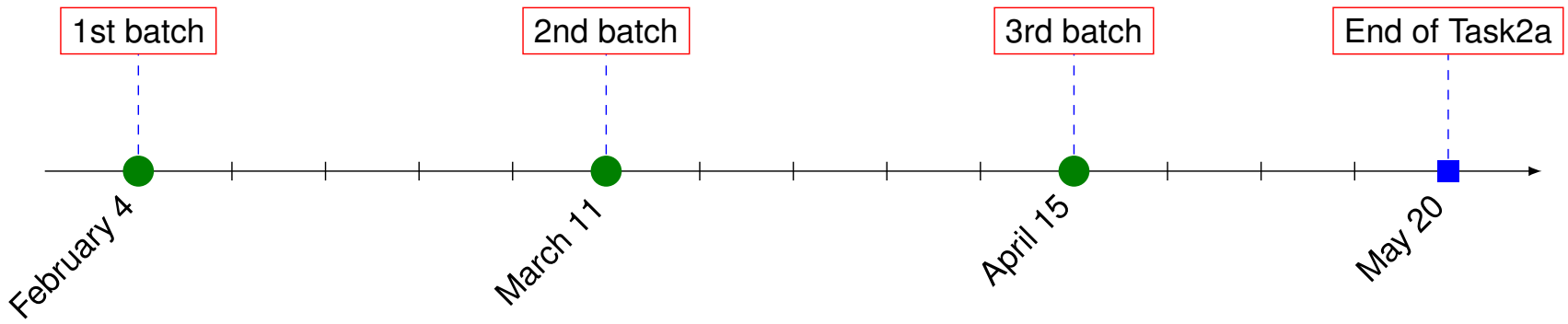
**REAL questions and answers**

**of many different types**

**created by bio-medical experts**

# Task 2A

## Schedule



# Task 2A

## Hierarchical text classification

Basic statistics about the **training** data

| | **version 2013** | **version 2014** | **version 2014 (2)** |
|---|---|---|---|
| **Articles** | 10,876,004 | 12,628,968 | 4,458,300 |
| **Total labels** | 26,563 | 26,831 | 26,631 |
| **Labels per article** | 12.55 | 12,72 | 13,20 |
| **Size in GB** | 18 | 20,31 | 6,4 |

Number of articles for each **test** dataset in each batch.

| **Week** | **Batch 1** | **Batch 2** | **Batch 3** |
|---|---|---|---|
| 1 | 4440 (3319) | 4085 (3422) | 4342 (3009) |
| 2 | 4721 (3734) | 3496 (2788) | 8840 (5883) |
| 3 | 4802 (3884) | 4524 (3274) | 3702 (2860) |
| 4 | 3579 (2431) | 5407 (3923) | 4726 (3252) |
| 5 | 5299 (3693) | 5454 (3666) | 4533 (3252) |
| Total | 22,841 (17,061) | 22,966 (17,073) | 26,143 (18,256) |

# Task 2A

## Evaluation Measures

### Flat measures

- Accuracy (Acc.)
- Example Based Precision (EBP)
- Example Based Recall (EBR)
- Example Based F-Measure (EBF)
- Macro Precision/Recall/F-Measure (MaP, MaR,MaF)
- Micro Precision/Recall/F-Measure (MiP,MIR,MiF)

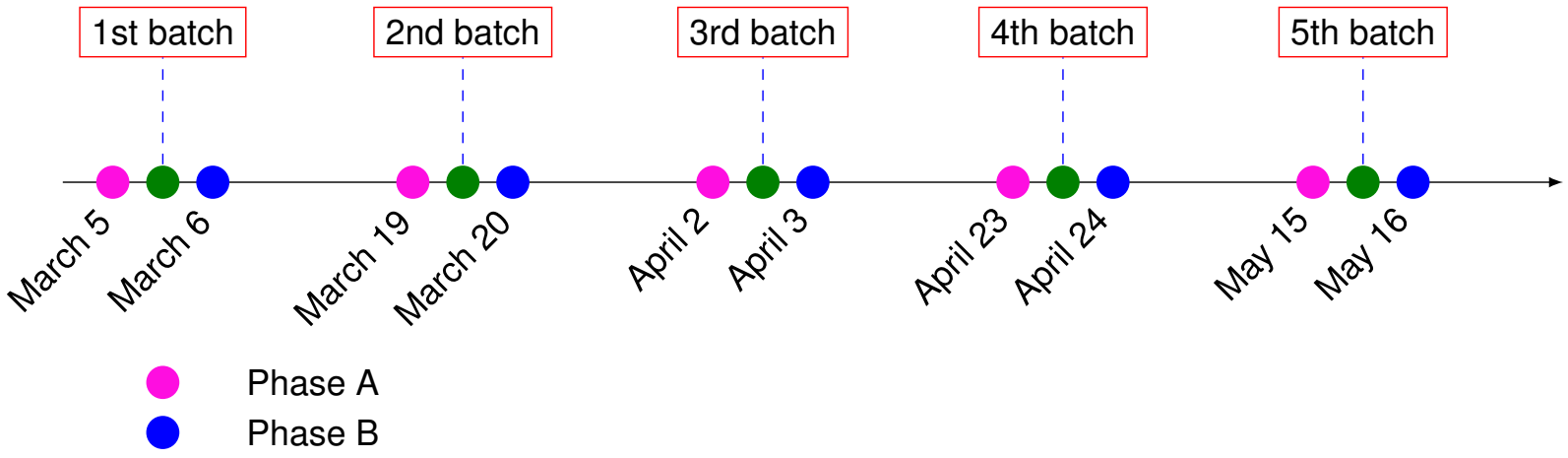
### Hierarchical measures

- Hierarchical Precision (HiP)
- Hierarchical Recall (HiR)
- Hierarchical F-Measure (HiF)
- Lowest Common Ancestor Precision (LCA-P)
- Lowest Common Ancestor Recall (LCA-R)
- Lowest Common Ancestor F-meData statistics for Task 1aasure (LCA-F)

A. Kosmopoulos, I. Partalas, E. Gaussier, G. Paliouras and I. Androutsopoulos: Evaluation Measures for Hierarchical Classification: a unified view and novel approaches. Data Mining and Knowledge Discovery (To appear)

# Task 2B

## Schedule



# Task 2B

## IR, QA, summarization

### Dataset

- 500 **Questions** and **gold reference answers** prepared by biomedical experts from around Europe.
  - Using tools/infrastructure developed by BioASQ.
- Four categories of questions:
  - Yes/No questions (both exact and ideal answer)
  - Factoids questions (both exact and ideal answer)
  - List questions (both exact and ideal answer)
  - Summary questions (ideal answer)

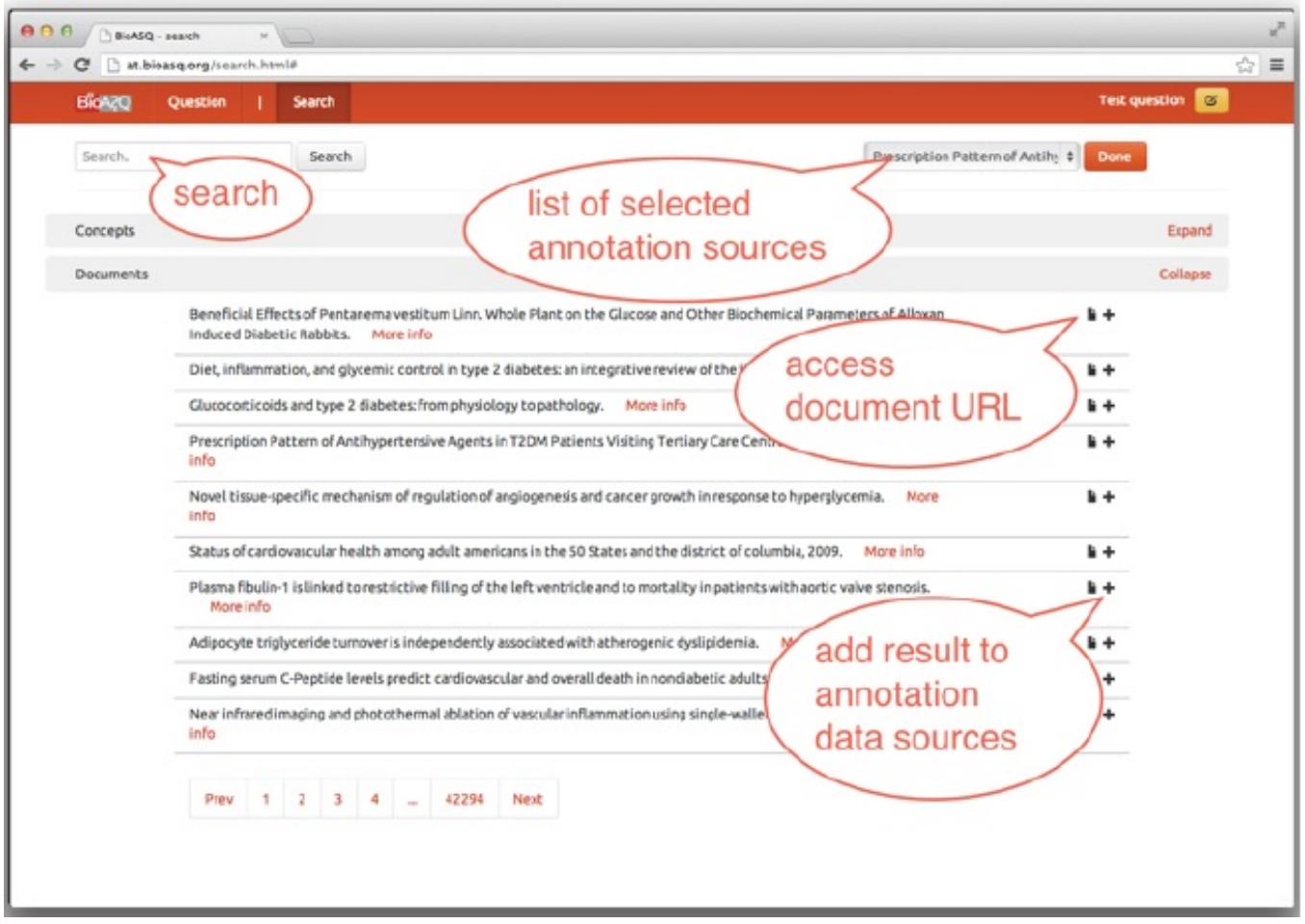
# Task 2B

## Examples of the different types of questions

- **Yes/No question:** *Is intense physical activity associated with longevity?*
- **Factoids question:** *Which is the protein (antigen) targeted by anti-Vel antibodies in the Vel blood group?*
- **List question:** *List the endoscopic diagnoses that have been reported in children with autism.*
- **Summary question:** *What is the role of thyroid hormone receptor alpha1 in insulin secretion?*

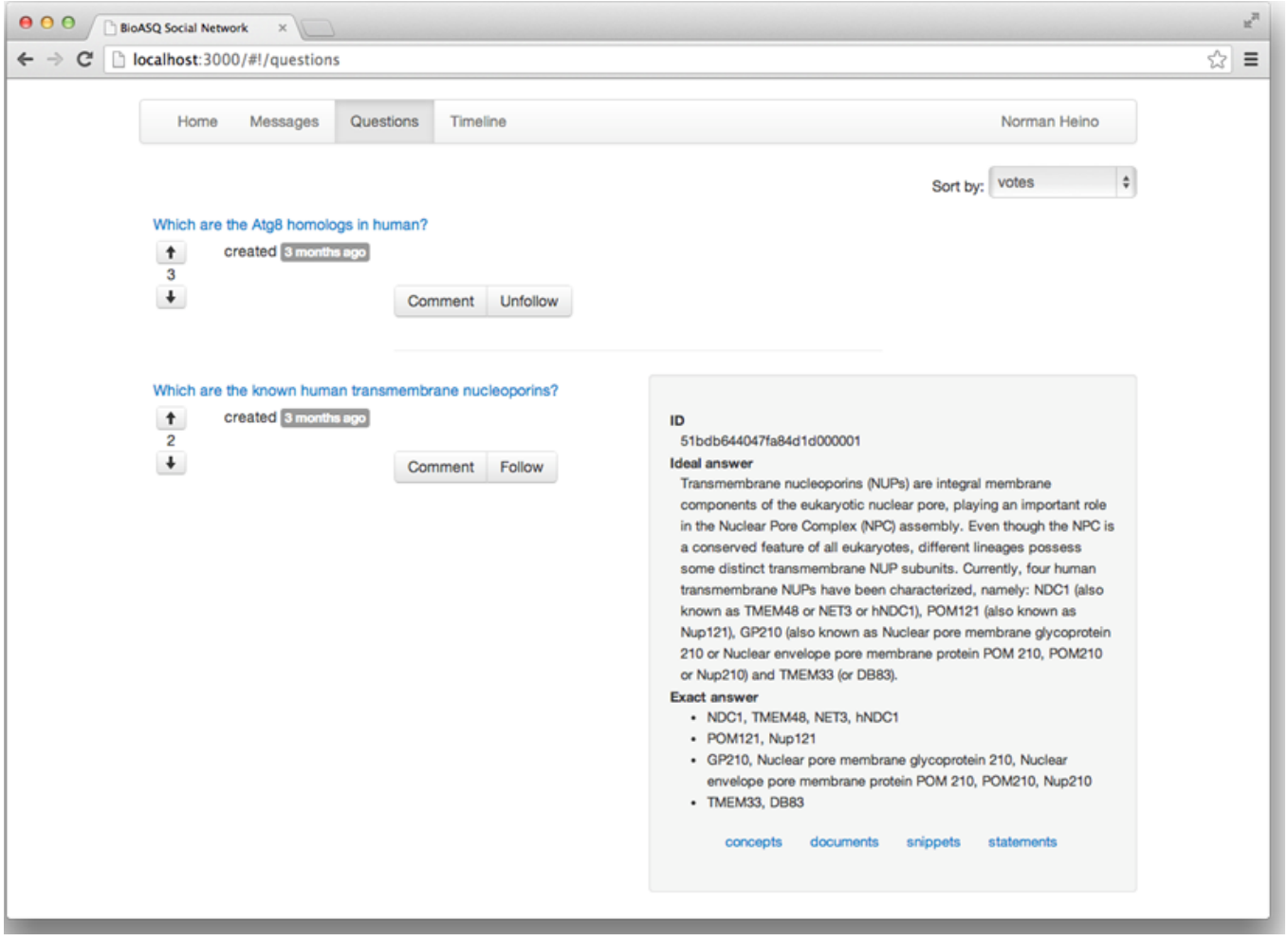
# Task 2B

**Annotation tool** for the creation of the data for QA



# Task 2B

## **Social network** to help extend data, and set up new challenges



# Task 2B

## Statistics on datasets

| Batch | Size | # of documents | # of snippets | # of concepts | # of triples |
|---|---|---|---|---|---|
| Training | 310 | 14.28 | 18.70 | 7.11 | 9.00 |
| Test 1 | 100 | 7.89 | 9.64 | 6.50 | 24.48 |
| Test 2 | 100 | 11.69 | 14.71 | 4.24 | 204.85 |
| Test 3 | 100 | 8.66 | 10.80 | 5.09 | 354.44 |
| Test 4 | 100 | 12.25 | 14.58 | 5.18 | 58.70 |
| Test 5 | 100 | 11.07 | 13.18 | 5.07 | 271.68 |
| **total** | 810 | 11.83 | 14.92 | 5.93 | 116.30 |

The numbers for the documents, snippets, concepts and triples refer to averages

# Task 2B

## Evaluation measures

- Evaluating **Phase A** (IR)

| Retrieved items | Unordered retrieval measures | Ordered retrieval measures |
|---|---|---|
| concepts | mean Precision, Recall, F-Measure | MAP, **GMAP** |
| articles | | |
| snippets | | |
| triples | | |

- Evaluating the **'exact' answers for Phase B** (Traditional QA)

| Question type | Participant response | Evaluation measures |
|---|---|---|
| yes/no | 'yes' or 'no' | **Accuracy** |
| factoid | up to 5 entity names | strict and lenient accuracy, **MRR** |
| list | a list of entity names | **mean** Precision, Recall, **F-measure** |

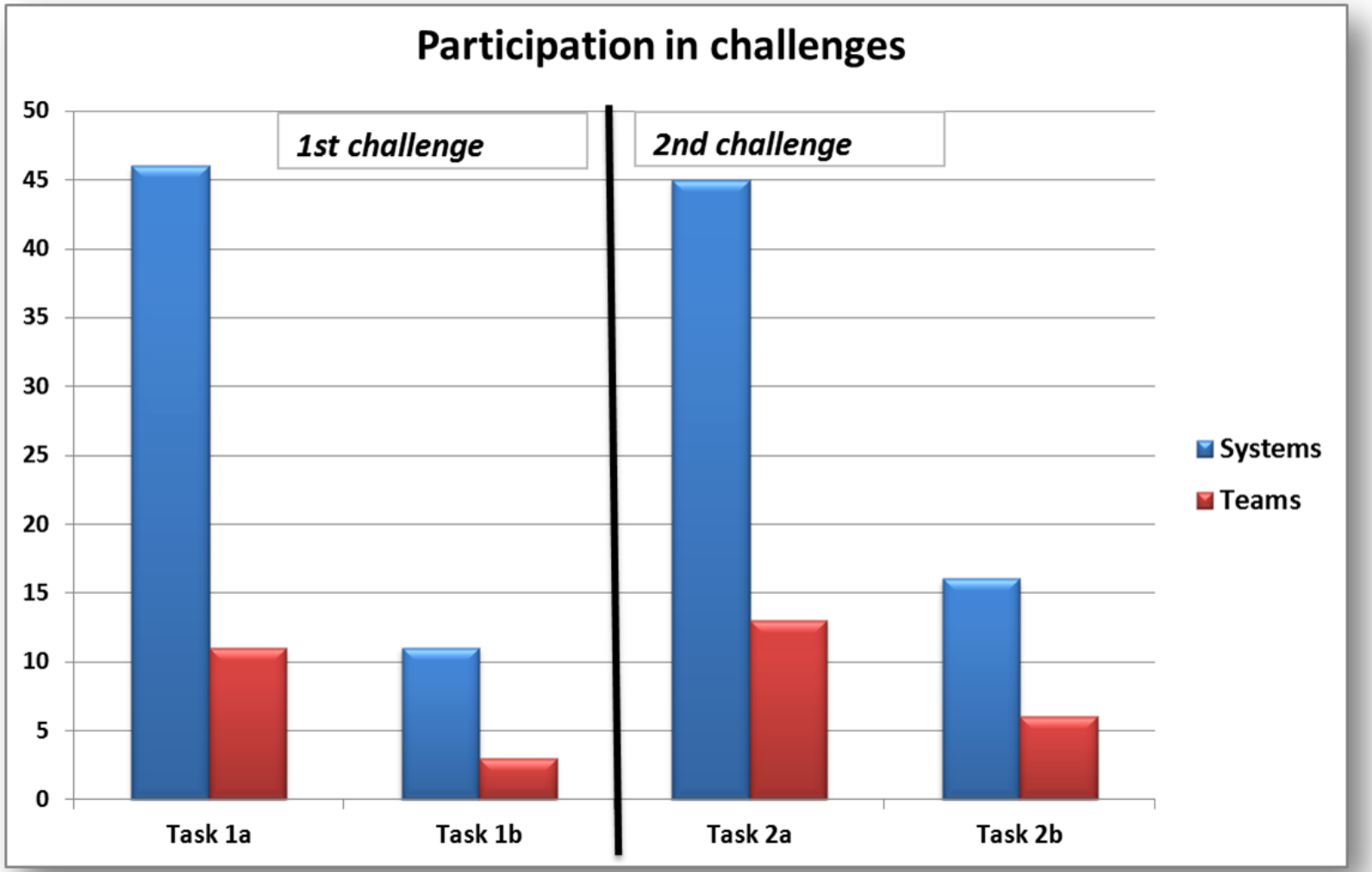
- Evaluating the **'ideal' answers for Phase B** (Query-focused Summarization)

| Question type | Participant response | Evaluation measures |
|---|---|---|
| any | paragraph-sized text | ROUGE-2, ROUGE-SU4, **manual scores*** (Readability, Recall, Precision, Repetition) |

*with the help of BioASQ Assessment tool.

# Challenge evaluation

## Comparison with first challenge participation

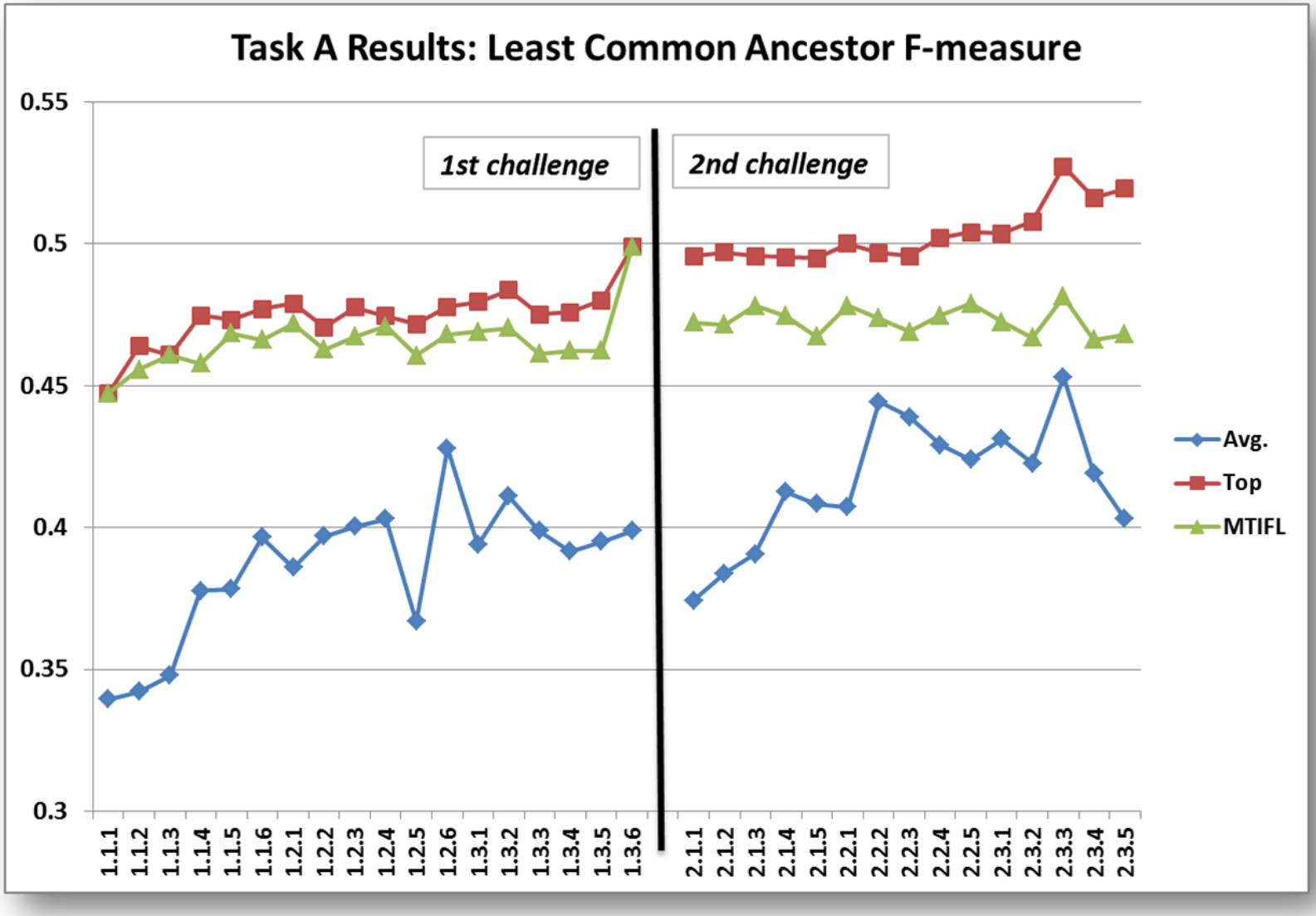


# Challenge evaluation

## Overall participation

# Challenge evaluation

## Results



# Conclusions and Perspectives

## What we've learnt

### Main conclusions

- Both tasks are challenging and interesting.
- It is difficult for humans to provide all required golden truth.
- Manual assessment and improvement of the data was necessary in task 2b.
- Evaluation is an open issue in both tasks.
- Satisfactory participation in the both BioASQ challenges.

### Goals and perspectives

- Continue after the end of the project:
  - Task 2a continues running in non-challenge mode.
  - Oracle for continuous testing has been announced.
  - Social network for data creation and challenge set-up.
  - BioASQ 3 will run next year.

# Stay Tuned!

## BioASQ project

**Visit www.bioasq.org**
**Follow @BioASQ**

**Call for paper:**
Journal of Bio-Medical Semantics
**Supplement on Semantics-Enabled Biomedical Information Retrieval**
Deadline: 30 November 2014

**Stay tuned for BioASQ 3**