Results of the fifth edition of the BioASQ Challenge

A. Nentidis, K. Bougiatiotis, A. Krithara, **G. Paliouras** and I. Kakadiaris

NCSR “Demokritos”, University of Houston

4th of August 2017

BioNLP Workshop, Vancouver

Introduction
What is BioASQ

A competition

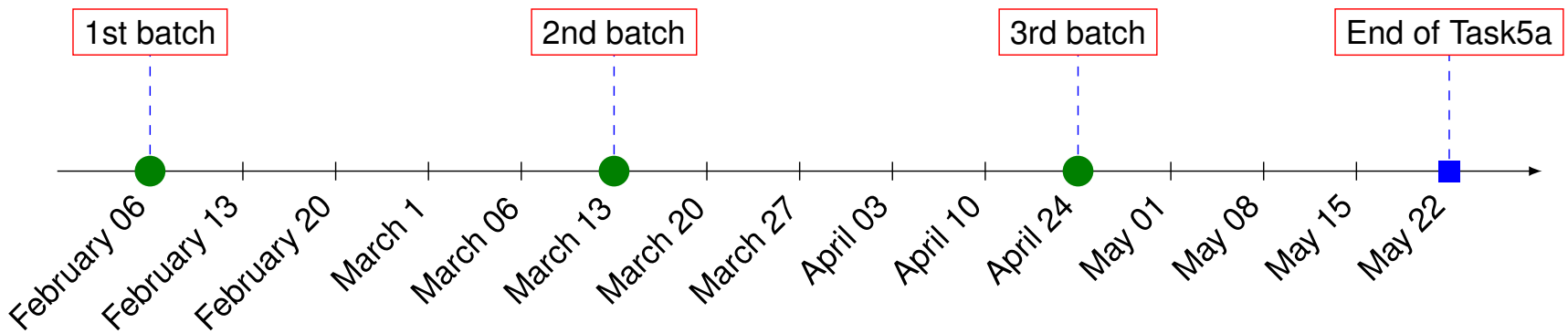
- BioASQ is a series of **challenges** on **biomedical semantic indexing** and **question answering (QA)**.
- Participants are required to semantically index content from **large-scale** biomedical resources (e.g. MEDLINE) and/or
- to assemble data from **multiple heterogeneous sources** (e.g. scientific 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 unclassified MEDLINE articles.**
- Participants have 21 hours to assign **MeSH terms** to the articles.
- **Evaluation** based on annotations of **MEDLINE curators.**

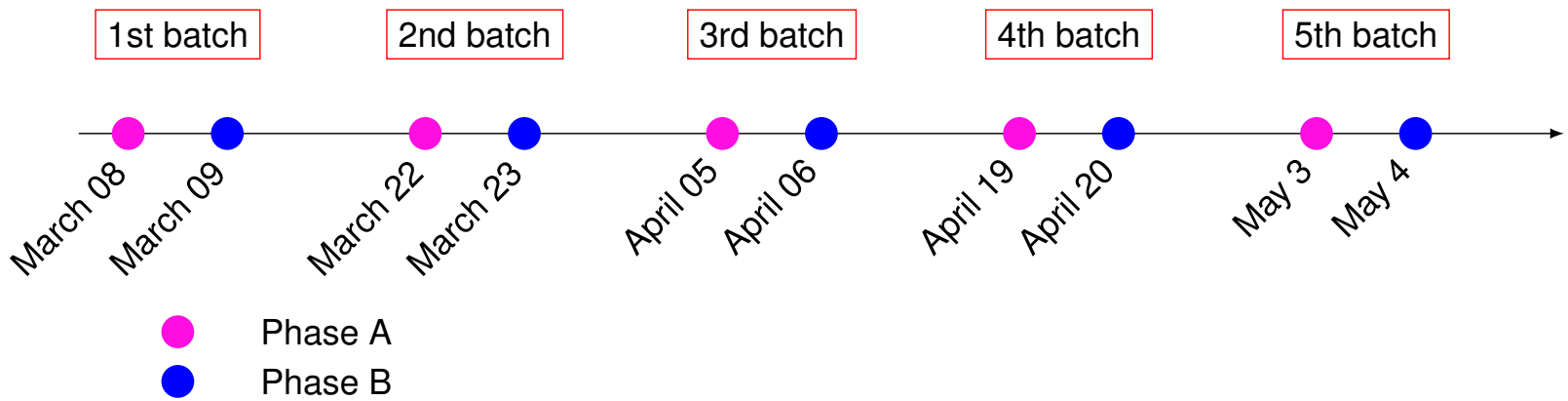


Presentation of the challenge

Tasks

Task B: IR, QA, summarization

- Organizers distribute **English biomedical questions.**
- Participants have 24 hours to provide: relevant **articles, snippets, concepts, triples, exact answers, ideal answers.**
- **Evaluation:** both **automatic** (GMAP, MRR, Rouge etc.) and **manual** (by biomedical experts).

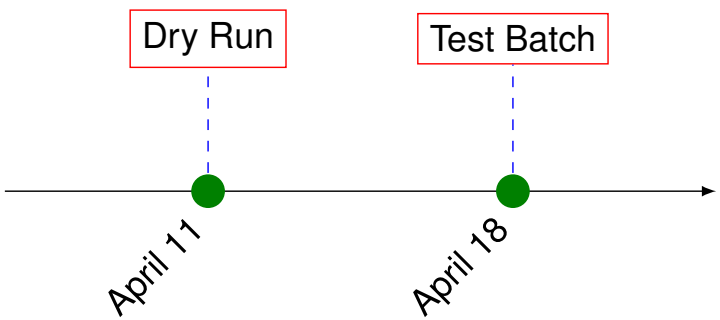


Presentation of the challenge

New task

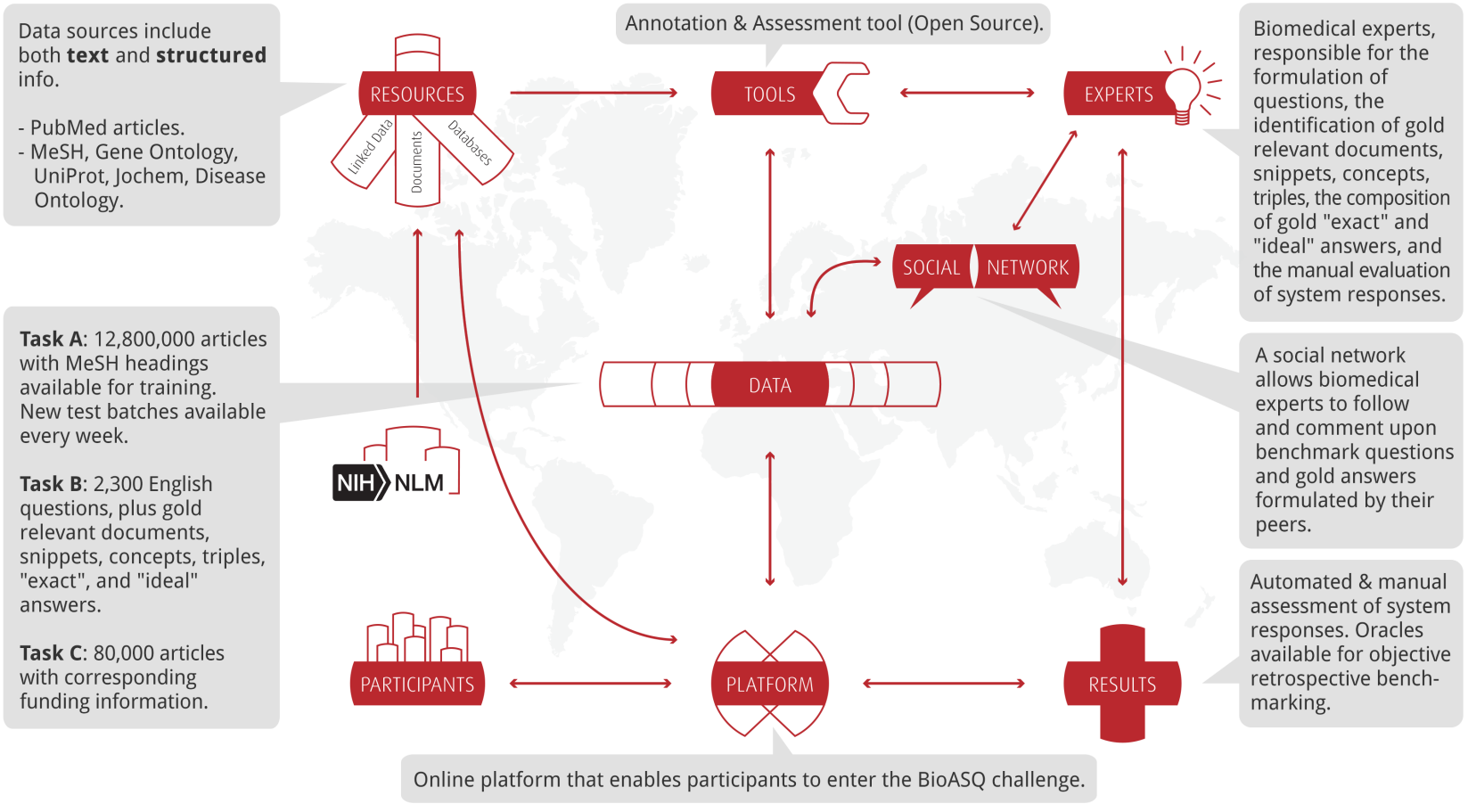
Task C: Funding Information Extraction

- Organizers distribute **PMC full-text articles.**
- Participants have 48 hours to extract: **grant-IDs, funding agencies, full grants** (i.e. the combination of a grant-ID and the corresponding funding agency).
- **Evaluation** based on annotations of **MEDLINE curators.**



Presentation of the challenge

BioASQ ecosystem



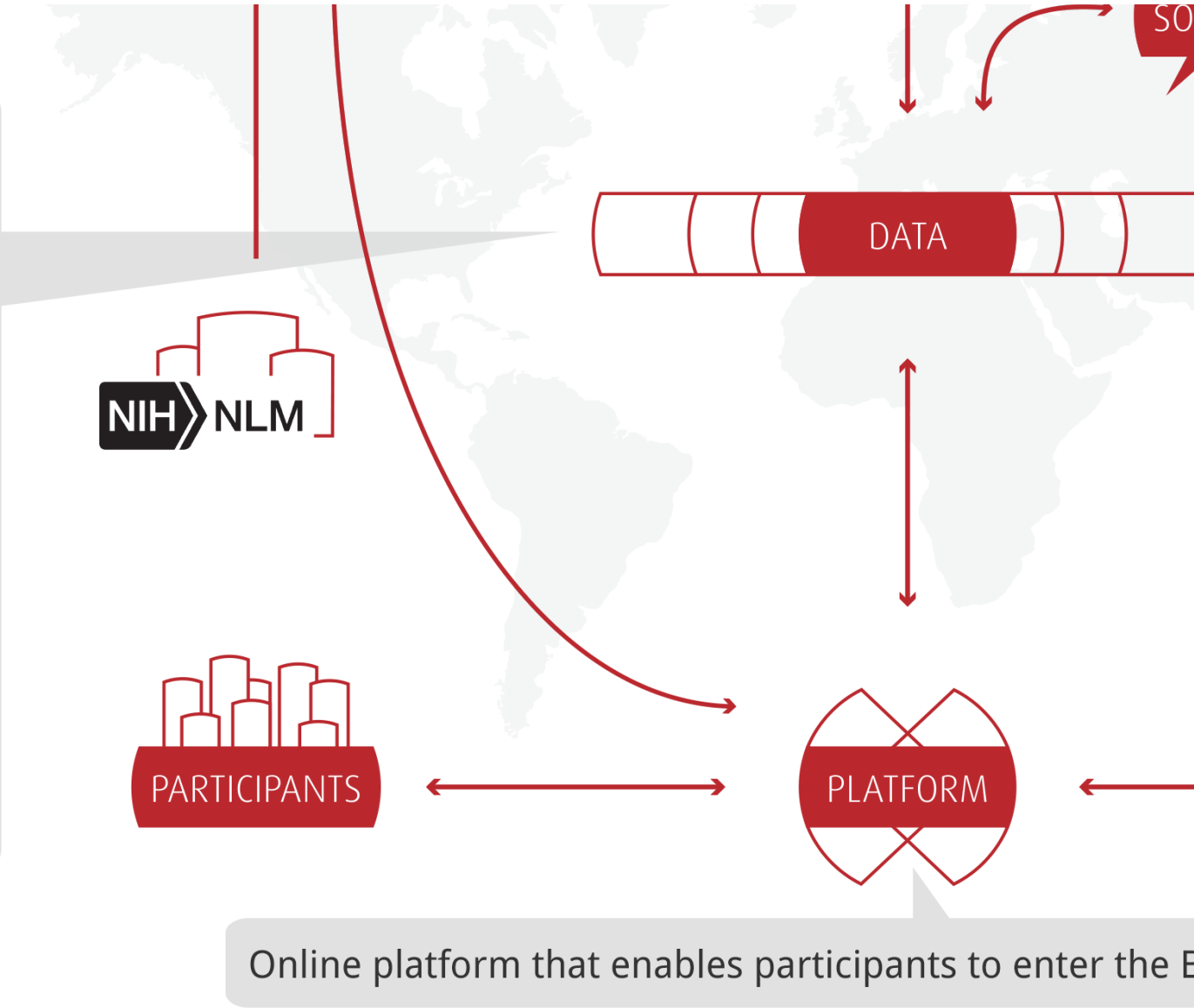
Presentation of the challenge

BioASQ ecosystem

Task A: 12,800,000 articles with MeSH headings available for training. New test batches available every week.

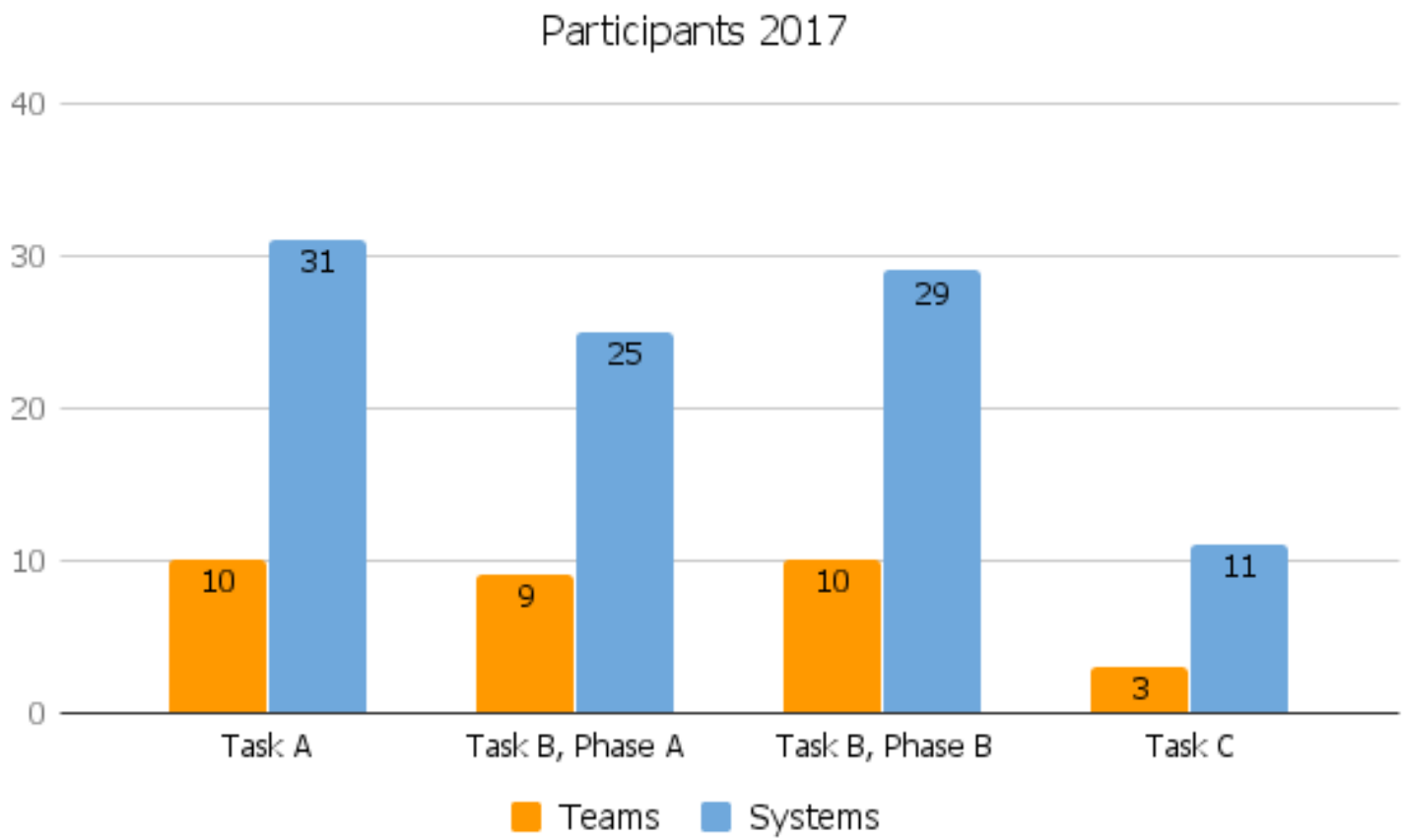
Task B: 2,300 English questions, plus gold relevant documents, snippets, concepts, triples, "exact", and "ideal" answers.

Task C: 80,000 articles with corresponding funding information.



Presentation of the challenge

Per task



Task 5A

Hierarchical text classification

- **Training** data

	version 2015	version 2016	version 2017
Articles	11,804,715	12,208,342	12,834,585
Total labels	27,097	27,301	27,773
Labels per article	12.61	12.62	12.66
Size in GB	19	19.4	20.5

- **Test** data

Week	Batch 1	Batch 2	Batch 3
1	6,880 (6,661)	7,431 (7,080)	9,233 (5,341)
2	7,457 (6,599)	6,746 (6,357)	7,816 (2,911)
3	10,319 (9,656)	5,944 (5,479)	7,206 (4,110)
4	7,523 (4,697)	6,986 (6,526)	7,955 (3,569)
5	7,940 (6,659)	6,055 (5,492)	10,225 (984)
Total	40,119 (34,272)	33,162 (30,934)	42,435 (21,323)

The numbers in parentheses are the annotated articles for each test dataset.

Task 5A

System approaches

- **Feature Extraction**: Representing each abstract
  - *tf-idf* of words and bi-words
  - *doc2vec* embeddings of paragraphs
- **Concept Matching**: Finding relevant MeSH labels
  - *k-NN* between article-vector representations
  - *Linear SVM* binary classifiers for each MESH label
  - *Recurrent Neural Networks* for sequence-to-sequence prediction
  - *UIMA-ConceptMapper* and *MeSHLabeler* tools for boosting NER and Entity-to-MeSH matching
  - *Latend Dirichlet Allocation and Labeled LDA* utilizing topics found in abstracts
  - *Ensemble* methodologies and stacking

Task 5A

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-measure (LCA-F)

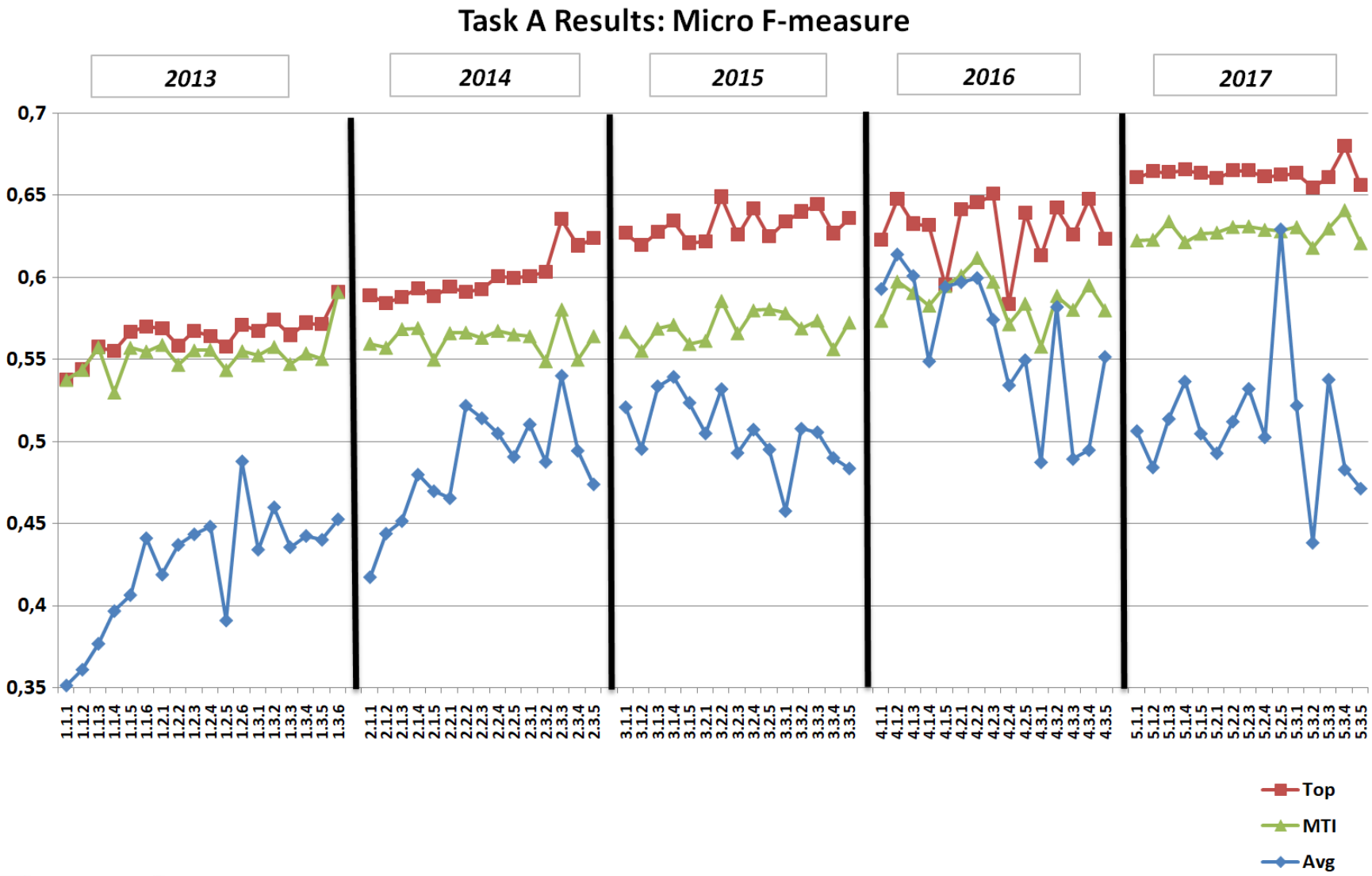
A. Kosmopoulos, I. Partalas, E. Gaussier, G. Paliouras and I. Androutsopoulos: Evaluation Measures for Hierarchical Classification: a unified view and novel approaches. Data Mining and Knowledge Discovery, 29:820-865, 2015.

Task 5A results

Evaluation

- Systems ranked using **MiF** (flat) and **LCA-F** (hierarchical).
- Results, in all batches and for both measures :
  1. **Fudan**
  2. **AUTH-Atypon**

Task 5A results



Task 5B

Statistics on datasets

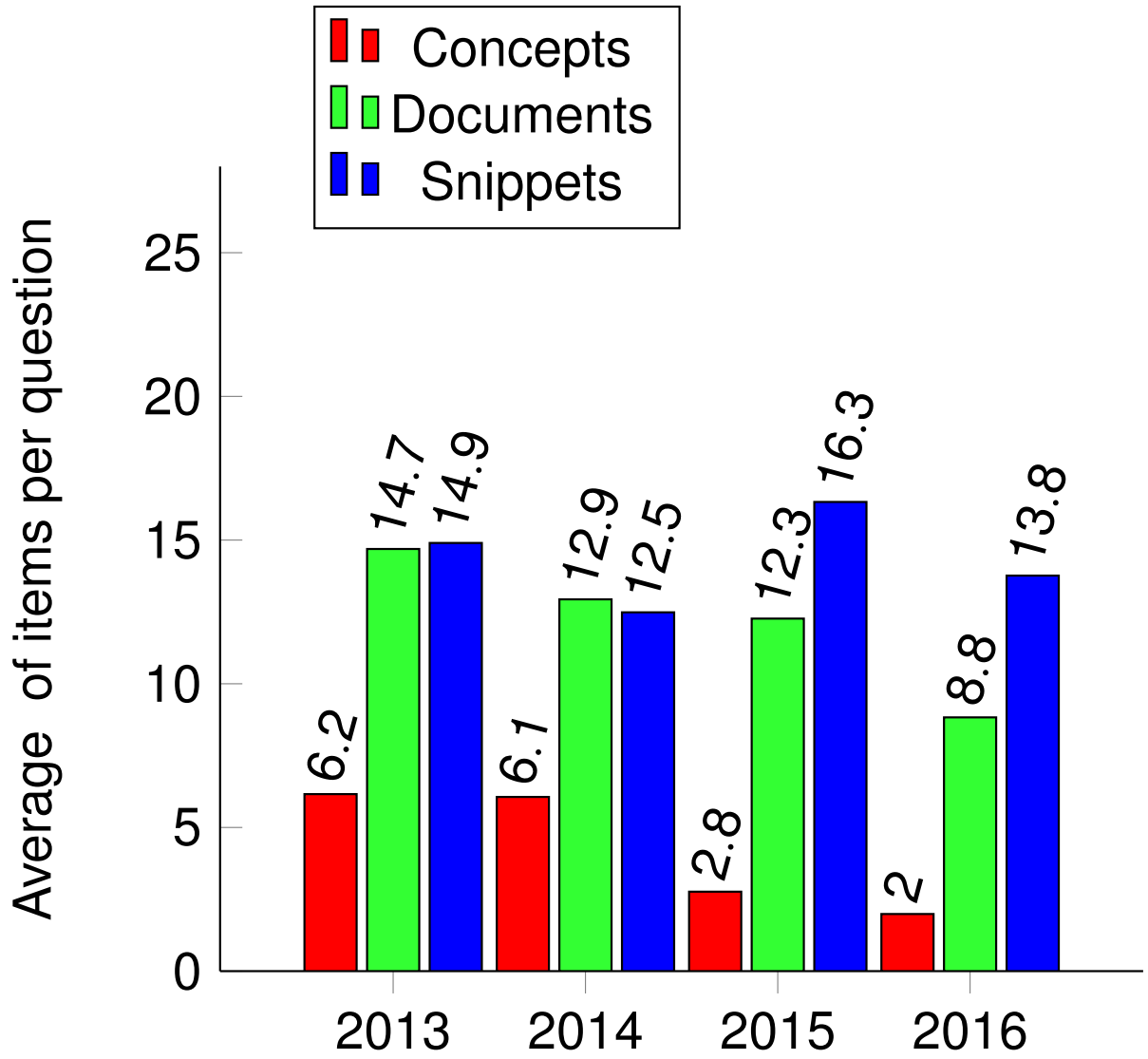
Batch	Size	# of documents	# of snippets
Training	1,799	11.86	20.38
Test 1	100	4.87	6.03
Test 2	100	3.49	5.13
Test 3	100	4.03	5.47
Test 4	100	3.23	4.52
Test 5	100	3.61	5.01
total	2,299		

The numbers for the documents and snippets refer to averages

Task 5B

Training Dataset Insights

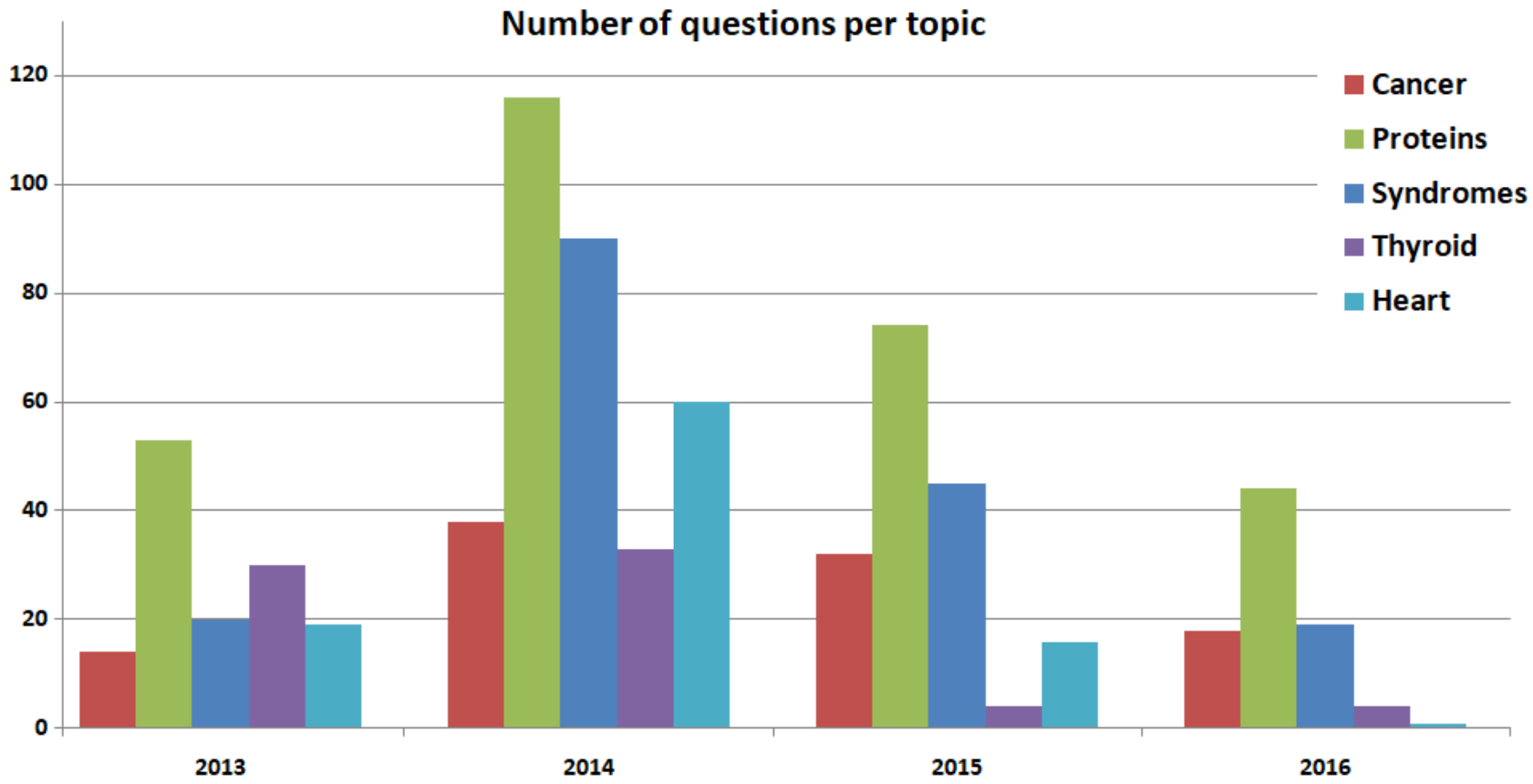
- **1799** Questions
  - 500 yes/no
  - 486 factoid
  - 413 list
  - 400 summary
- **13** Experts
- ≈ **3450** unique biomedical concepts



Task 5B

Training Dataset Insights

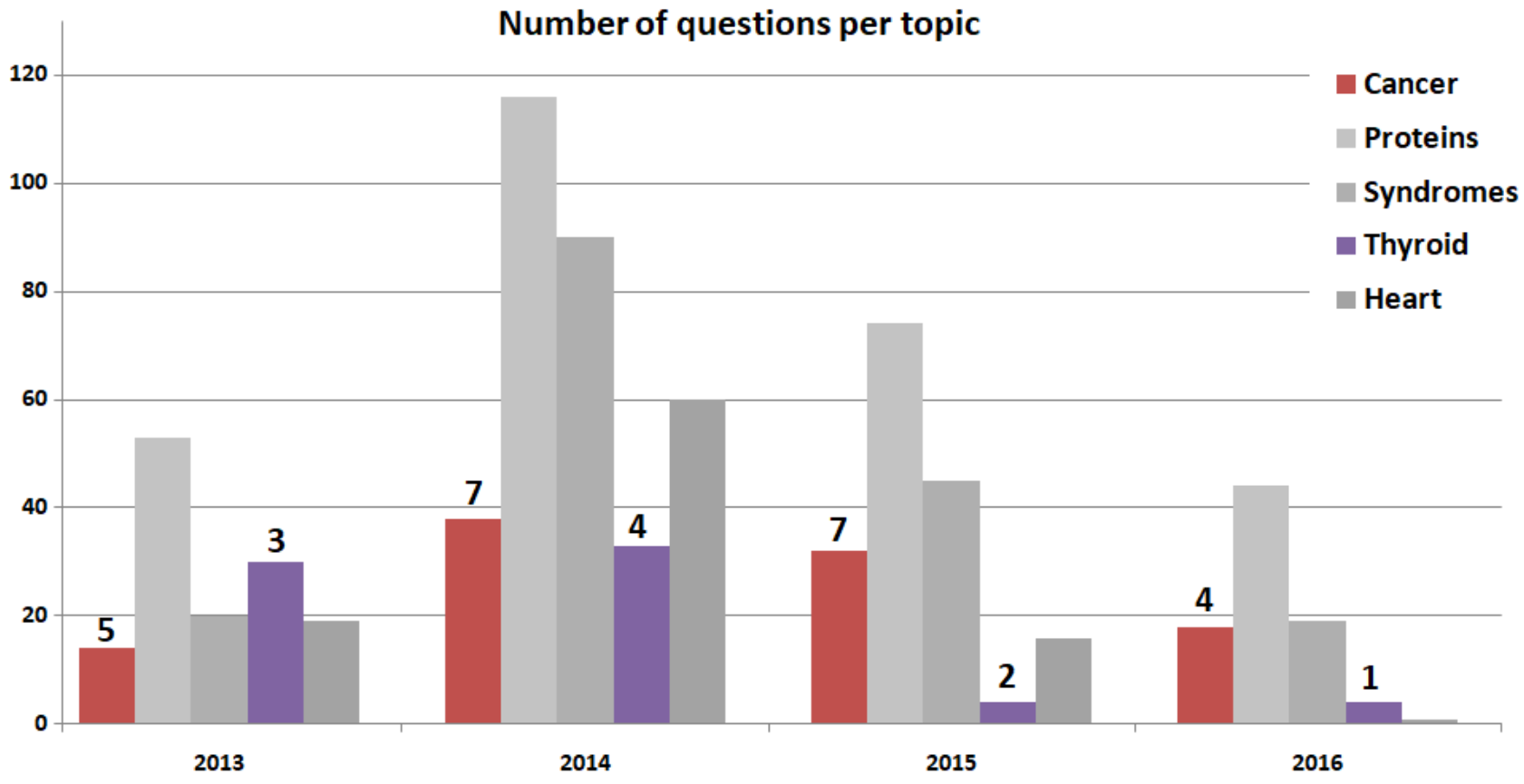
- Broad terms (e.g. proteins, syndromes)
- More specific terms (e.g. cancer, heart, thyroid)



Task 5B

Training Dataset Insights

- Number of questions related to **cancer** vs **thyroid** per year
- The numbers on top of the bars denote the contributing experts



Task 5B

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.

Task 5B

System approaches

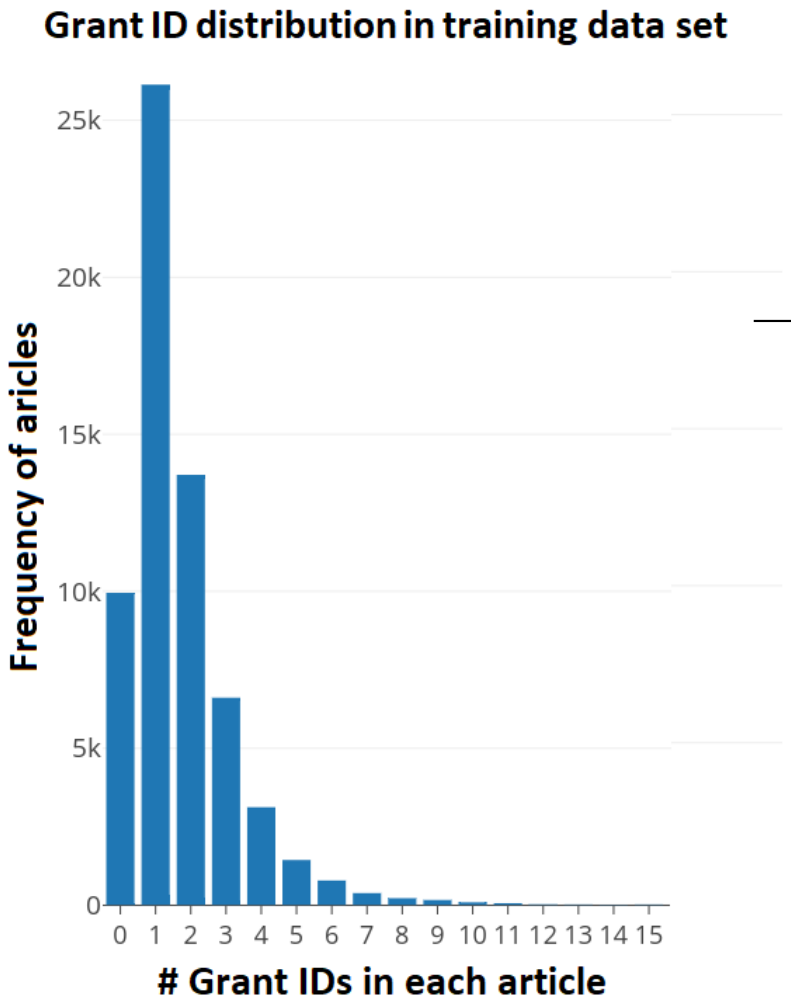
- **Question analysis**: Rule-based, regular expressions, ClearNLP, Semantic role labeling (SRL), Stanford Parser, tf-idf, SVD, word embeddings.
- **Query expansion**: MetaMap, UMLS, sequential dependence models, ensembles, LingPipe.
- **Document retrieval**: BM25, UMLS, SAP HANA database, Bag of Concepts (BoC), statistical language model.
- **Snippet selection**: Agglomerative Clustering, Maximum Marginal Relevance, tf-idf, word embeddings.
- **Exact answer generation**: Standford POS, PubTator, FastQA, SQuAD, Semantic role labeling (SRL), word frequencies, word embeddings, dictionaries, UMLS.
- **Ideal answer generation**: Deep learning (LSTM, CNN, RNN), neural nets, Support Vector Regression.
- **Answer ranking**: Word frequencies.

Task 5B Results

- Our experts are currently assessing systems' responses
- The results will be announced in autumn

Task 5C

Statistics on datasets



	Training	Test
Articles	62,952	22,610
Grant IDs	111,528	42,711
Agencies	128,329	47,266
Time Period	2005-13	2015-17

- **104** unique agencies
- **92,437** unique grant IDs

Task 5C

Statistics on datasets

Task 5C

Evaluation measures

- A **subset** of the Grant IDs and Agencies mentioned in full text are available in ground truth data$\Rightarrow$ **Micro-Recall**
  - Each Grant ID (or lone Agency) must exist verbatim in the text
- Different scores for each subtask:
  - Grant IDs
  - Agencies
  - Full Grants

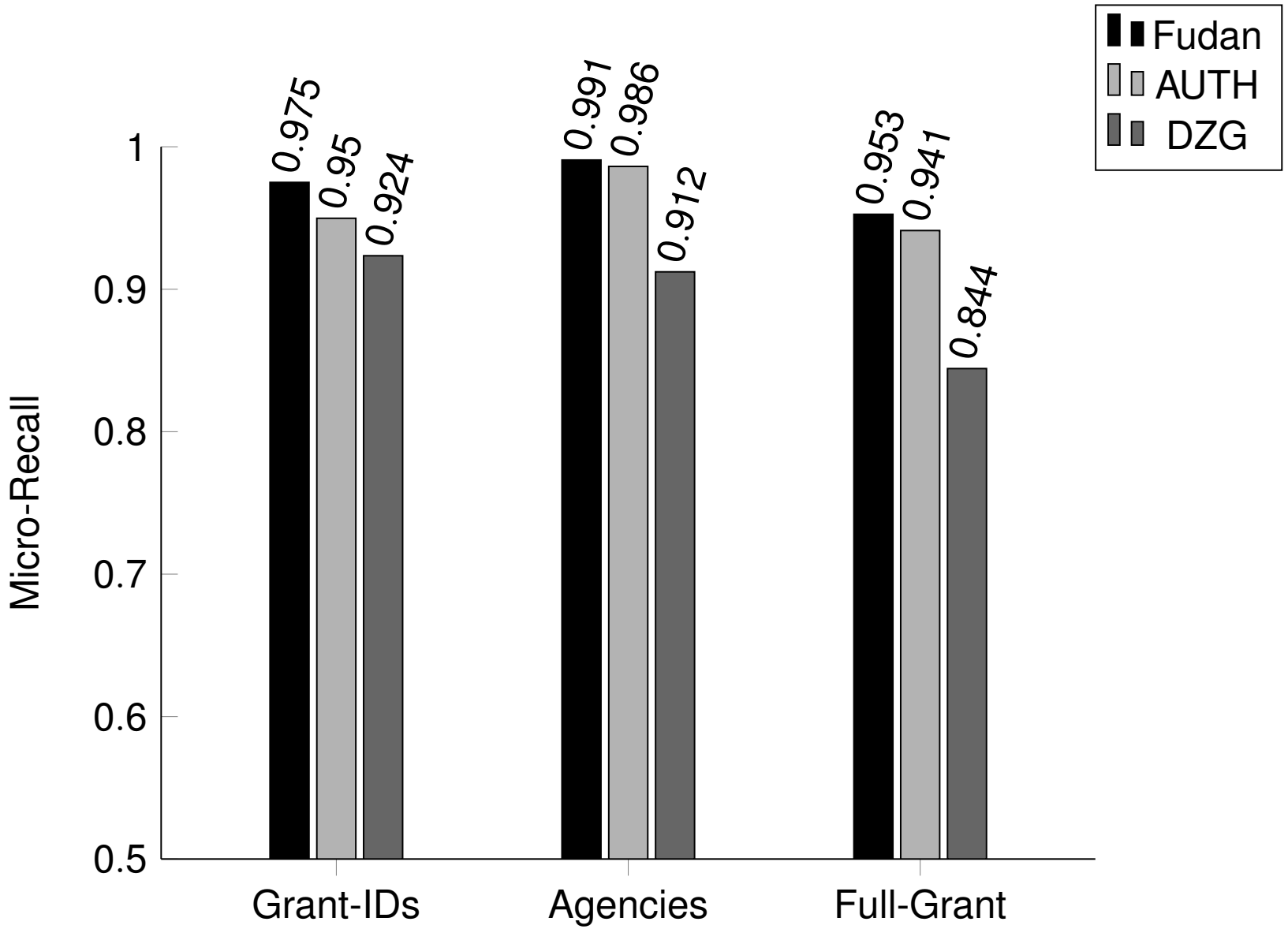
Task 5C

System approaches

- **Grant Support Sentences**: Identifying sentences containing grant information
  - Features: *tf-idf* of n-grams
  - Techniques: *SVM* and *Naive Bayes* for scoring, specific XML fields considered
- **Grant Information Extraction**: Detecing Grant-IDs and Agencies
  - Manually crafted *Regular Expressions*
  - *Heuristic Rules*
  - *Sequential Learning Models*, such as *Conditional Random Fields*, *Hidden Markov Models*, *Max Entropy Models*
  - Ensemble of classifiers for pairing Grant-IDs to Agencies

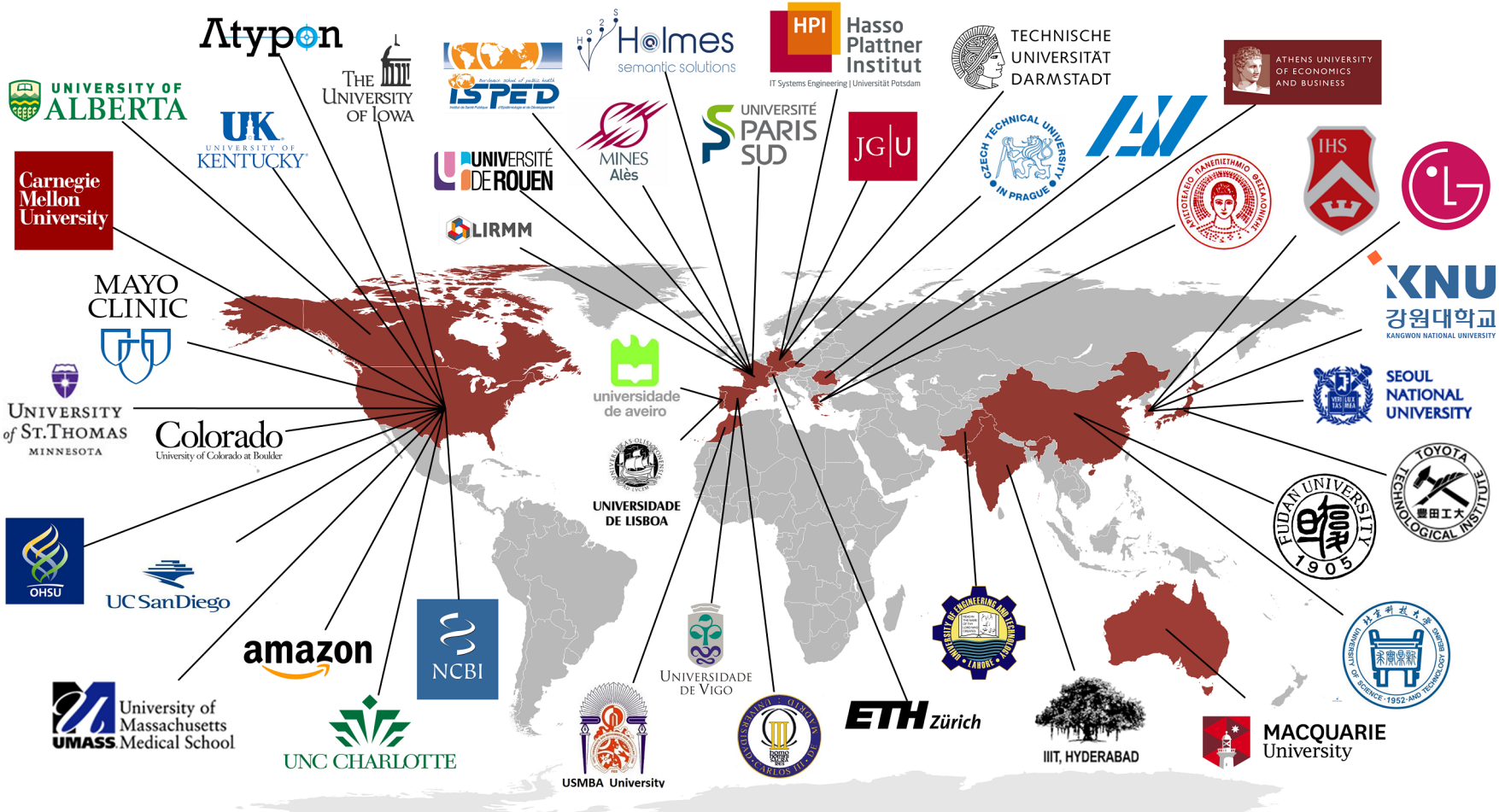
Task 5C

Results



Challenge Participation

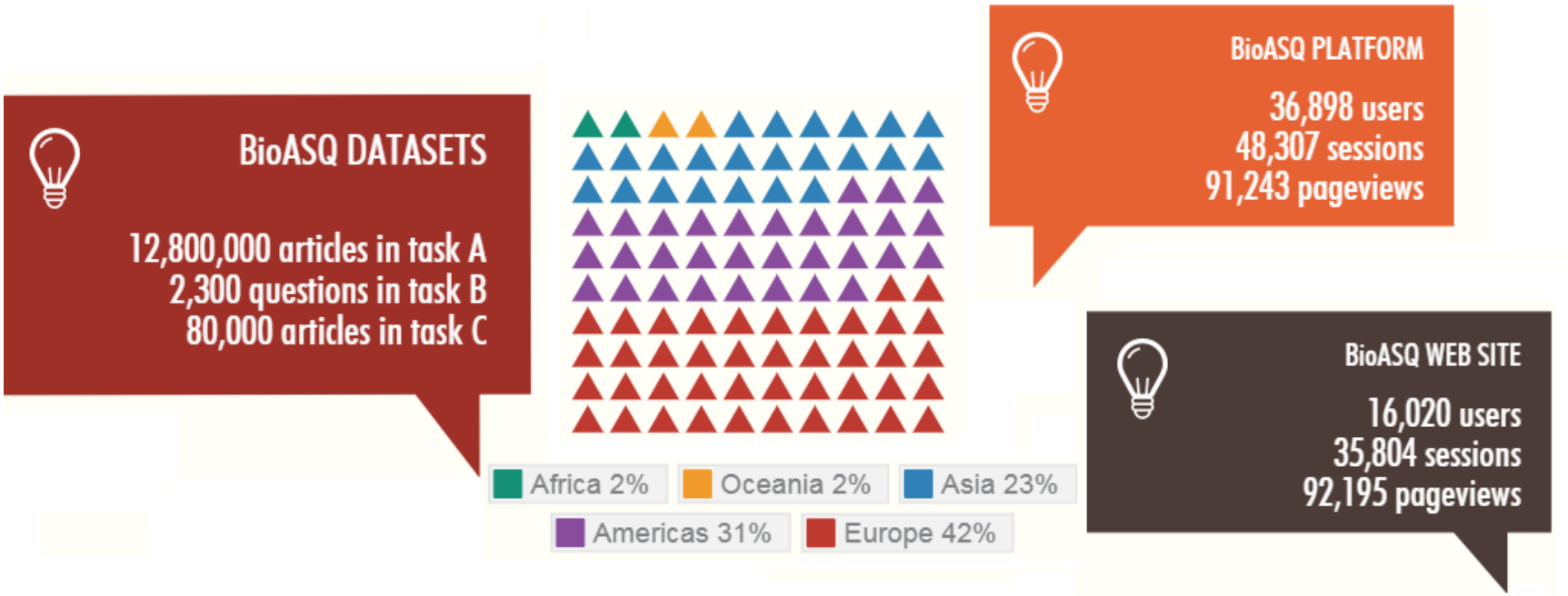
Overall



Conclusions and Prespectives

Goals and perspectives

- BioASQ will run in 2018.
- Continuous development of benchmark datasets.



Conclusions and Prespectives

Oracle for **continuous testing**

Task: Task A ▾ — Select the task you are submitting results for.

Test: Task 4a: Test batch 3, Week 5 ▾ — Specify the test set by choosing one from the drop down menu. The tests sets for both tasks can be downloaded from here and are those that been already used for the BioASQ challenge.

Your system: bioasq_baseline ▾ — Select one of your systems that will be used in the "Oracle Results" tab.

Your system results: Επιλογή αρχείου Δεν επιλέχθηκε κανένα αρχείο. — Select a file to upload that contains a JSON string with the answers of a test. The format of the JSON is described in the online guidelines of each task, e.g. here.

Submit

Attention: Calculating the evaluation results takes several minutes. Please, do not refresh the content.

Results

Annotated documents: 627 out of 3130.

Please, take a look at the results below and fill the following form:

Keep my results visible: ☑ — If enabled, your uploaded results will be visible in the oracle to any registered user. Otherwise, it will be visible only to you.

Save my score: ☑ — If enabled, it will replace the previous score for the selected system and testset in the BioASQ database.

Submit

Flat Measures

System	MiF	Acc.	EBP	EBR	EBF	MaP	MaR	MaF	MiP	MiR
auth1	0.5954	0.4247	0.5887	0.6133	0.5793	0.5659	0.4776	0.4593	0.5948	0.5959
Current Submission	0.5817	0.4091	0.5843	0.5994	0.5641	0.5481	0.4821	0.4634	0.5794	0.5841
d33p	0.5746	0.3978	0.6150	0.5473	0.5507	0.5626	0.3897	0.3811	0.6143	0.5397
Default MTI	0.5854	0.4165	0.6036	0.5934	0.5711	0.5369	0.5173	0.4960	0.5967	0.5745

Collaborations

- **NLM**
  - Task A design and baselines
  - Task C design and baselines
- **CMU**
  - OAQA Baselines for task B
- **DBCLS**
  - BioASQ and PubAnnotation : Using linked annotations in biomedical question answering (BLAH3)
- **iASiS**
  - Question answering over big heterogeneous biomedical data for precision medicine

Grateful to the BioASQ consortium

BioASQ started as a European FP7 project, with the following partners:

- National Centre for Scientific Research "Demokritos" (GR)
- Transinsight GmbH (DE)
- Universite Joseph Fourier (FR)
- University Leipzig (DE)
- Universite Pierre et Marie Curie Paris 6 (FR)
- Athens University of Economics and Business Research Centre (GR)

Sponsors

PLATINUM SPONSOR

SILVER SPONSOR

Stay Tuned!

Visit www.bioasq.org
Follow @BioASQ

BioASQ 6 to be announced soon!