

# Semantic Web Challenge on Tabular Data to KG Matching

**Kavitha Srinivas**, IBM Research, USA

**Ernesto Jiménez-Ruiz**, City, University of London, UK

**Oktie Hassanzadeh**, IBM Research, USA

**Jiaoyan Chen**, University of Oxford, UK

**Vasilis Efthymiou**, IBM Research, USA

# Motivation

- **Tabular data** in the form of **CSV files** is the common input format in a data analytics pipeline.
- Gaining **semantic understanding** will be very valuable for data integration, data cleaning, data mining, machine learning and knowledge discovery tasks.
- Lack of a **systematic evaluation** framework for Semantic Web solutions.

## Adding Semantics to Tabular Data: Challenge Tasks

- Assigning a semantic type (e.g., a KG class) to an (entity) column (**CTA task**)
- Matching a cell to a KG entity (**CEA task**)
- Assigning a KG property to the relationship between two columns (**CPA task**)

*(\*) We assume the existence of a (possibly incomplete) **Knowledge Graph (KG)** relevant to the domain.*

*(\*\*) We relied on DBpedia KG.*

# Adding Semantics to Tabular Data: Example

|     |                  |                |                  |            |
|-----|------------------|----------------|------------------|------------|
|     | Countries        | has population | Cities           |            |
| 1   | China            | 1,377,516,162  | Beijing          | 09-22-2016 |
| 2   | India            | 1,291,999,508  | New Delhi        | 09-22-2016 |
| 3   | United States    | 323,990,000    | Washington, D.C. | 09-22-2016 |
| 4   | Indonesia        | 258,705,000    | Jakarta          | 07-01-2016 |
| 5   | Brazil           | 206,162,929    | Brasilia         | 09-22-2016 |
| ... |                  |                |                  |            |
| 16  | Congo            | 82,310,000     | Kinshasa         | 07-01-2016 |
| ... |                  |                |                  |            |
| 26  | Burma            | 54,363,426     | Naypyidaw        | 07-01-2016 |
| ... |                  |                |                  |            |
| 122 | Congo            | 4,741,000      | Brazzaville      | 07-01-2016 |
| ... |                  |                |                  |            |
| 194 | Falkland Islands | 2,563          | Stanley          | 04-15-2012 |

Republic of the Congo

Democratic Republic of the Congo

(\*) Adapted from Efthymiou et al. Matching Web Tables with Knowledge Base Entities: From Entity Lookups to Entity Embeddings. ISWC 2017

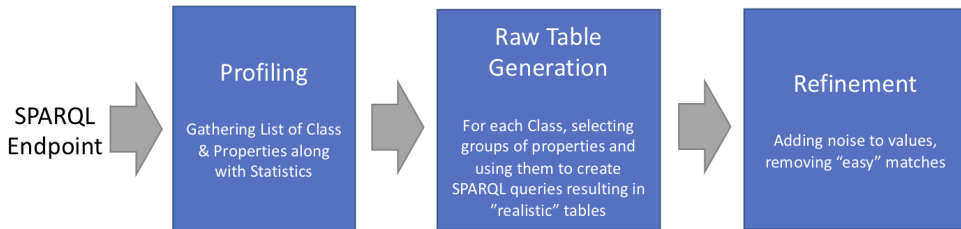
# Datasets

- **Round 1** (sandbox): extended **T2Dv2 dataset**
- **Round 2** (fine-tuning): **Wikipedia tables** dataset + **automatically generated** dataset
- **Round 3** (limited tests): **automatically generated** dataset
- **Round 4** (limited tests): **automatically generated** dataset with only **hard cases**

Tables and ground truth for all rounds are made publicly available at:

<https://doi.org/10.5281/zenodo.3518539>

# Automatic Dataset Generator

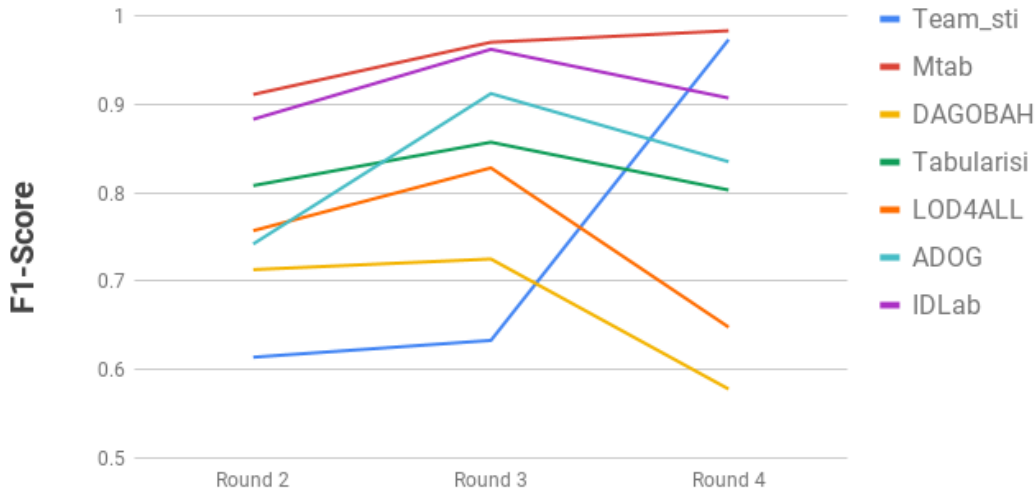


# Participation

- 7 systems stable across tasks and rounds
- Good starting to create community

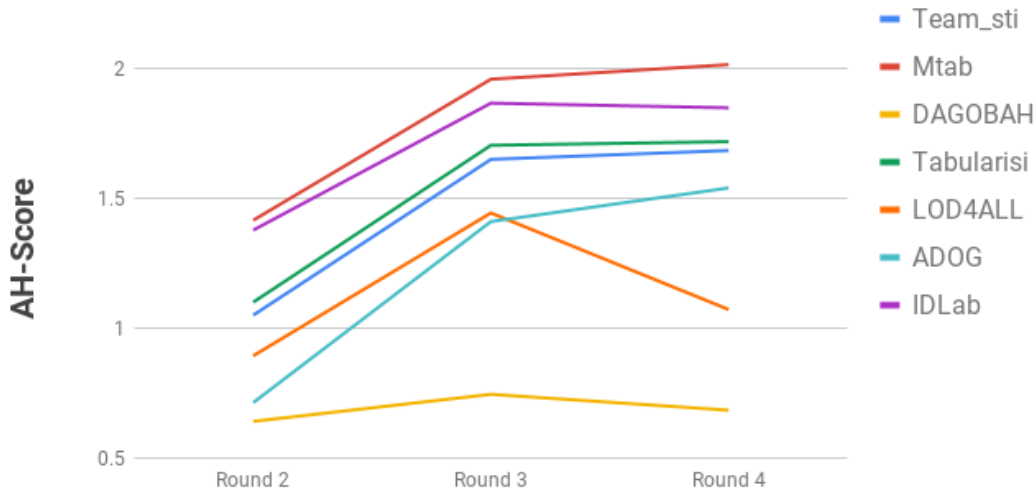
| #                   | Round 1 | Round 2 | Round 3 | Round 4 |
|---------------------|---------|---------|---------|---------|
| <b>Participants</b> | 17      | 11      | 9       | 8       |
| <b>CTA</b>          | 13      | 9       | 8       | 7       |
| <b>CEA</b>          | 11      | 10      | 8       | 8       |
| <b>CPA</b>          | 5       | 7       | 7       | 7       |

## CEA Task

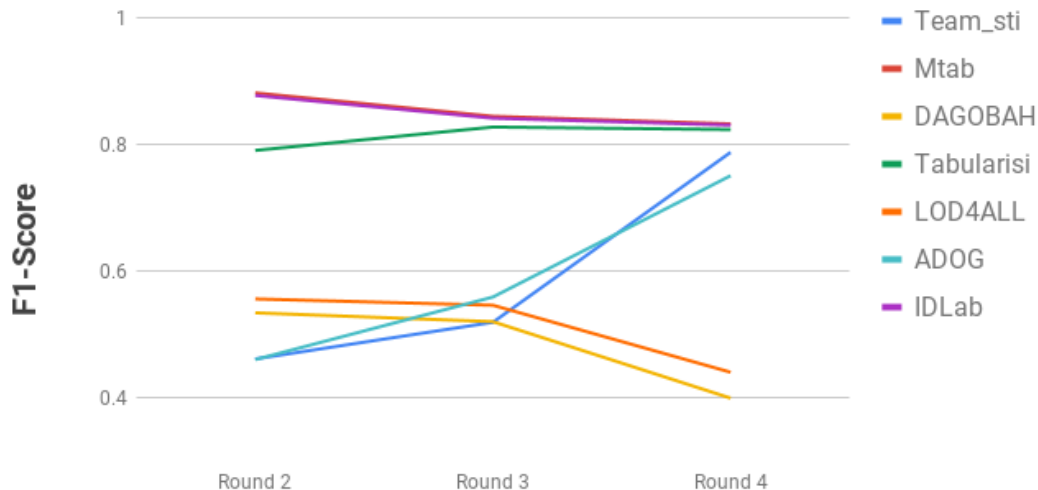




## CTA Task



## CPA Task



# Acknowledgements

- All participants
- Challenge organisers and their institutions
- AICrowd and Arjun Nemani
- Our sponsors: IBM Research and SIRIUS
- ISWC and OM organisers

IBM Research



DEPARTMENT OF  
COMPUTER  
SCIENCE

The  
Alan Turing  
Institute



CITY UNIVERSITY  
LONDON

Auckland 2019  
ISWC

# ISWC Challenge Prizes

- Prizes sponsored by **IBM Research** and **SIRIUS**  
(Norwegian Center for Scalable Data Access):  
<http://www.sirius-labs.no/>

**IBM Research**

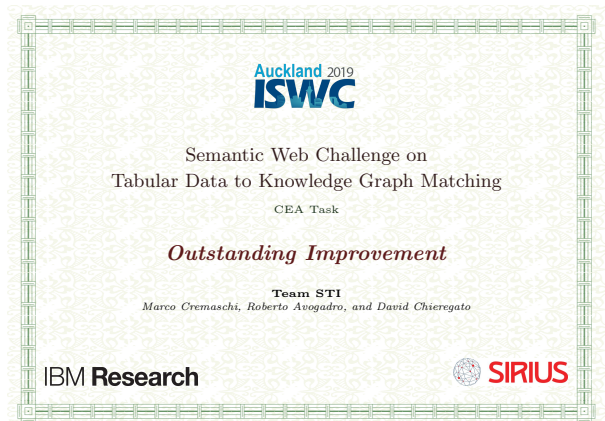


**SIRIUS**

# CEA Outstanding Improvement

# CEA Outstanding Improvement

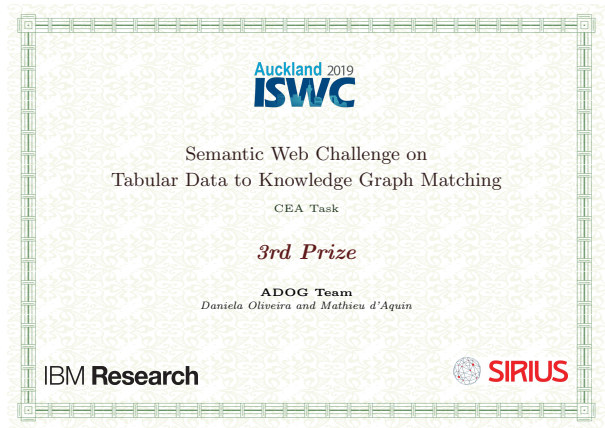
- **TEAM\_STI**: Marco Cremaschi, Roberto Avogadro, and David Chierigato



# CEA 3rd Prize

# CEA 3rd Prize

– **ADOG**: Daniela Oliveira and Mathieu d'Aquin

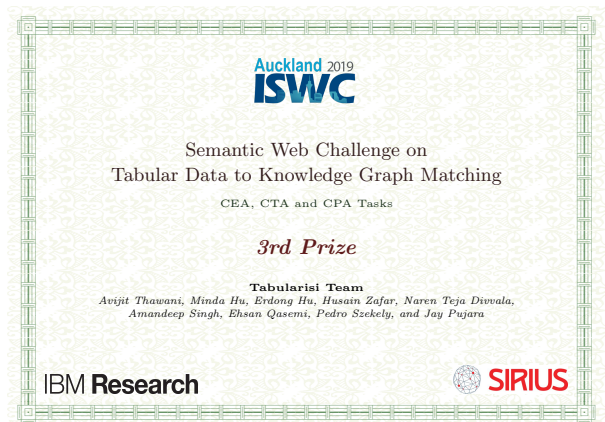




# Global 3rd Prize

## Global 3rd Prize

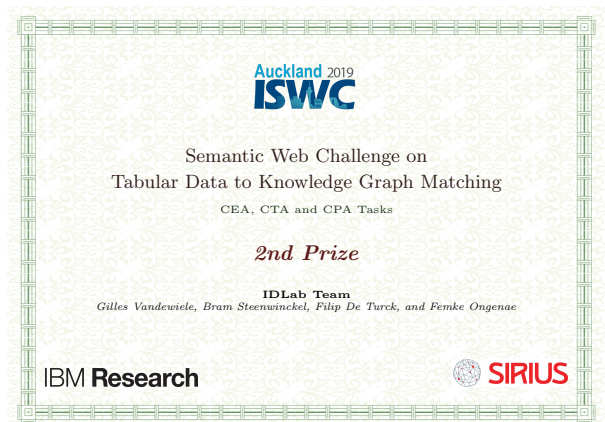
- **Tabularisi**: Avijit Thawani, Minda Hu, Erdong Hu, Husain Zafar, Naren Divvala, Amandeep Singh, Ehsan Qasemi, Pedro Szekely, Jay Pujara.



# Global 2nd Prize

## Global 2nd Prize

- **IDLab**: Gilles Vandewiele, Bram Steenwinckel, Filip De Turck, and Femke Ongenae.



# Global 1st Prize

# Global 1st Prize

- **MTab**: Phuc Nguyen, Natthawut Kertkeidkachorn, Ryutaro Ichise and Hideaki Takeda.



# Challenge Talks

8 min + 1/2 questions:

- MTab
- Tabularisi
- Team STI
- Team DAGOBAAH