**Discovery Approach Description**

The description provided below will be used to evaluate the approach developed by your team to automatically identify public web sources of annual financial data of MNE Groups. This description will be evaluated by the Evaluation panel based on the criteria described in the Evaluation tab of the Discovery Challenge and used for the ranking of your team for the Reusability and Innovativeness Awards.

# **Methodology *(Data-driven approaches; Availability and quality of documentation)***

Please provide a detailed description of the methodology used for automatically identifying the public web sources of annual financial data of MNE Groups. The description should contain (1) the data processing steps, (2) the methods and models used, (3) references to the scientific papers/sources that present the methods and models used, and (4) the time it took to process the data set.

Bear in mind that the workflow will be also evaluated based on the criteria for the Reusability and Innovativeness Awards.

*This section will be evaluated for:*

*(1) Data-driven approaches: The described approach is evaluated based on whether it is data-driven rather than heuristic. More data-driven approaches will receive higher scores. The code will be inspected visually by the evaluation panel. Well documented code which allows evaluators to determine the steps of the approach will yield higher scores.*

*(2) Availability and quality of documentation: Based on the clarity of the provided documentation describing the approach, the evaluation panel is asked to assess the likeliness that the described approach can successfully reproduce the solution submitted by the team for the Accuracy award.*

|  |
| --- |
| **Todo: Spiegare l’approccio di Raffaele** |

## **Architecture *(Algorithm reusability and scalability)***

Please provide a description of the architecture of your approach. A diagram of the architecture is considered of additional value. Indicate what modifications would be required to apply the approach to similar datasets on a larger scale.

*This section will be evaluated for:*

1. *The described approach is evaluated based on:*
   1. *Are the components well separated and encapsulated to allow for independent modification?*
   2. *What is the degree of modification required to add new companies?*
   3. *Is it possible to implement the approach by running it on parallel machines?*
   4. *What is the extent of effort required to track performance?*
   5. *Is the method robust if scaled to larger number of cases?*
   6. *Is the performance of the approach stable as scale increases?*

|  |
| --- |
|  |

# **Hardware Specifications *(Algorithm reusability and scalability)***

Please describe the hardware specifications of the machines that were used to run the methodology.

*This section will be evaluated for:*

1. *Algorithm reusability and scalability*
   1. *Is it possible to implement the approach by running it on parallel machines?*
   2. *Is the method robust if scaled to larger number of cases?*
   3. *Is the performance of the approach stable as scale increases?*

**Machine 1**

|  |  |
| --- | --- |
| CPUs | CPU name and capacity |
| GPUs | GPU name and capacity |
| TPUs | TPU name and capacity |
| Disk space | The space required to calculate and store the data |

**Machine 2**

|  |  |
| --- | --- |
| CPUs | CPU name and capacity |
| GPUs | GPU name and capacity |
| TPUs | TPU name and capacity |
| Disk space | The space required to calculate and store the data |

# **Libraries**

Please provide the libraries used for approach, if any, as well as the links to these libraries, if available.

|  |
| --- |
|  |

## **Similarities/differences to State-of-the-Art techniques *(Originality of the approach)***

Please provide a list of similarities and differences between the used methodology and to the state-of-the-art techniques.

*This section will be evaluated for:*

*(1) the Originality of the approach criterion: compare the approach used to the state of the art, i.e. currently published approaches that are closest to the approach applied for the submission, and the extent to which the submission represents an improvement over these approaches. The submission will be evaluated based on the degree to which it is new or unique. This could be in terms of technology, methodology, or application.*

|  |
| --- |
|  |

## **Contribution to scientific field *(Future orientation)***

Please describe how your submission contributed to the scientific field, what impact it could have and what could potentially be future work to improve the solution.

*This section will be evaluated for:*

*(1) the Future orientation and impact criterion: The potential effect of the approach used will be evaluated. This includes the scale of impact it has on the problem of discovering sources of financial data from the Internet. The impact will be evaluated based on potential efficiency improvements and cost reductions.*

|  |
| --- |
|  |

## **Lessons Learned *(Future orientation)***

Please state any lessons learned during the competition.

*This section will be evaluated for:*

*(1) the Future orientation and impact criterion: what were the lessons learnt during the competition, and what could potentially be future work to improve the solution.*

|  |
| --- |
|  |

# **Short description of the Team – area of expertise**

Please provide a description of the team, your area of expertise and contact information.

|  |
| --- |
| **Paolo Barba: Data Scientist,** [**paolobarba50@gmail.com**](mailto:paolobarba50@gmail.com)**., cell. +39 3451600069**  **Raffaele Liguori: Data Engineer**  **Francesco Cellitti: Data Analyst**  **Gianmario Sangiovanni: phd Candidate Data Analysis** |