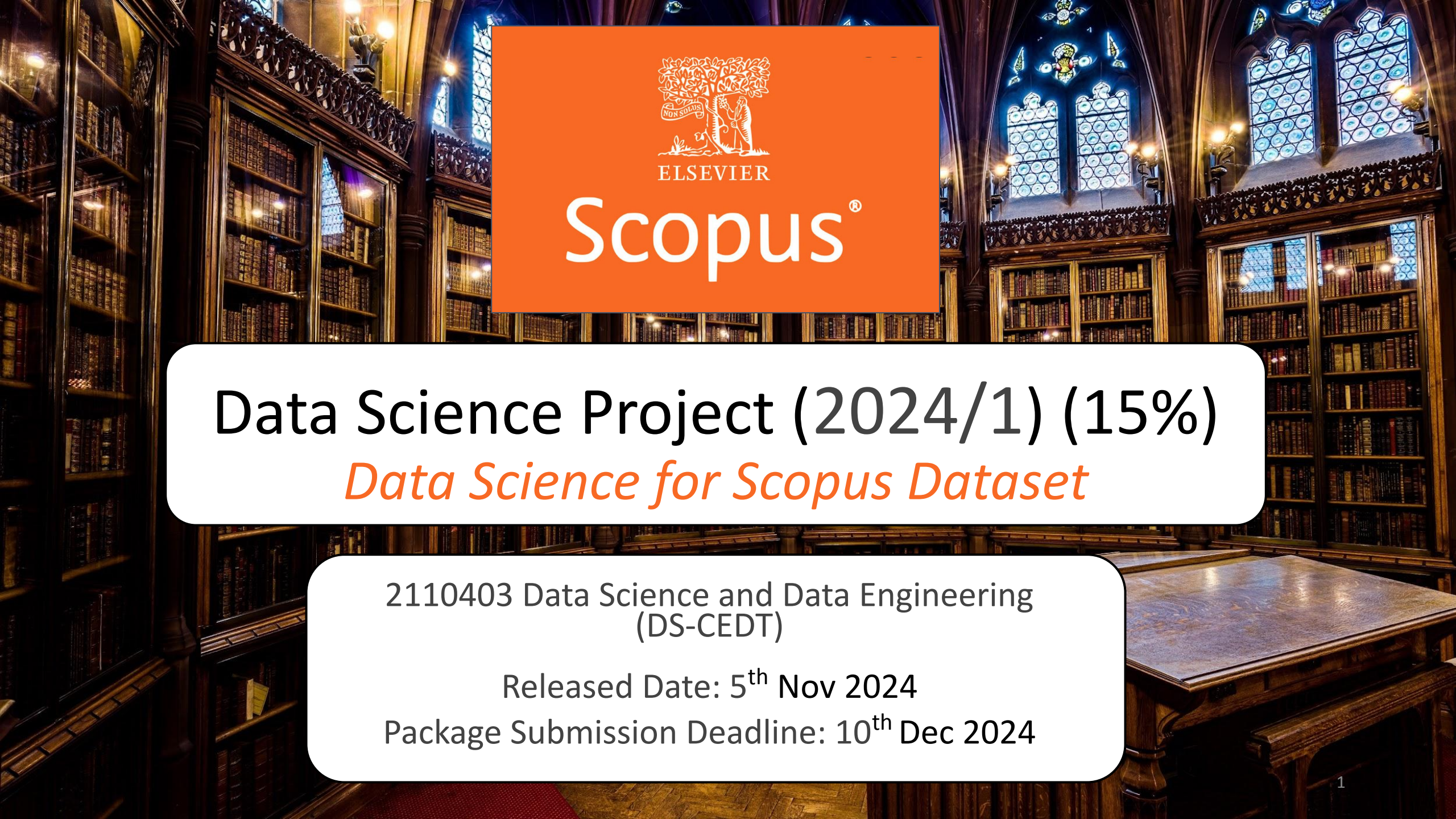


Data Science Project (2024/1) (15%)

Data Science for Scopus Dataset

2110403 Data Science and Data Engineering (DS-CEDT)

Released Date: 5th Nov 2024

Package Submission Deadline: 10th Dec 2024

Full raw Scopus data

- With the support of the CU Office of Academic Resources (2018 - 2023),
- Metadata of literatures in the Engineering field are provided
- The data contains documents in JSON format
- Example of useful data
  - Title and abstract
  - Document classification codes
  - Date of publication
  - Affiliations
  - Citations / references
  - Keywords

https://drive.google.com/file/d/107WikNVtve-QY7I7-pMsdFFHpAnNFxmO/view?usp=sharing

Objective

- Data science is a discipline aimed at data analysis, involving various components, e.g., AI/ML, data preparation, data engineering, data visualization, ML operations, and more.
- Therefore, the objective of this project is to build a practical pipeline and demonstrate diverse, **actionable (end-to-end)** data analysis ideas. Examples include:
  - Example 1: Data Science Pipeline including web scraping → Kafka → workflow control with Airflow → visualization with Power BI
  - Example 2: Big Data Analytics Pipeline including large data ingestion → processing with Spark → visualization through Power BI → storytelling insights

Project criteria

- Each group can have up to 6 members (maximum).
- The project should be a fully functional, end-to-end pipeline that demonstrates practical applications or yields insightful findings. The project must include at least the following 3 components:
  - Component 1: At least one AI/ML component.
  - Component 2: At least one Data Engineering (DE) component.
  - Component 3: At least one visualization (Viz) component that includes either geospatial analysis or graph visualization.
- Including more than three components will enhance the project's depth and interest, with the potential for additional points.

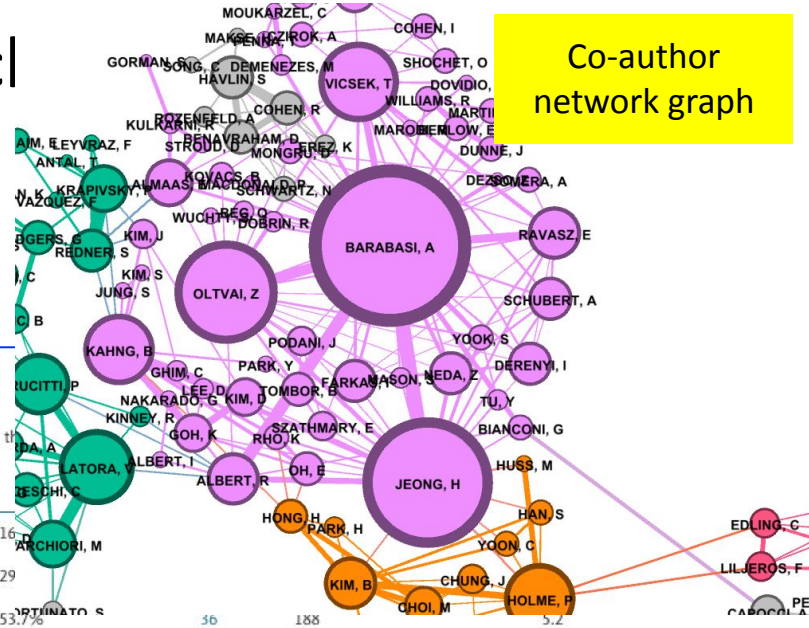
Data criteria

- The primary data requirement is to focus on analyzing the provided data from Scopus.
- Each group must also use web scraping or an API to add at least 1,000 additional papers from external sources, such as Scopus, DBLP, or Google Scholar.
  - This data collection is a separate task and does not count as part of the DE module.
- Incorporating additional data sources, such as author locations or affiliation coordinates, will enhance the project's appeal and may earn extra points.

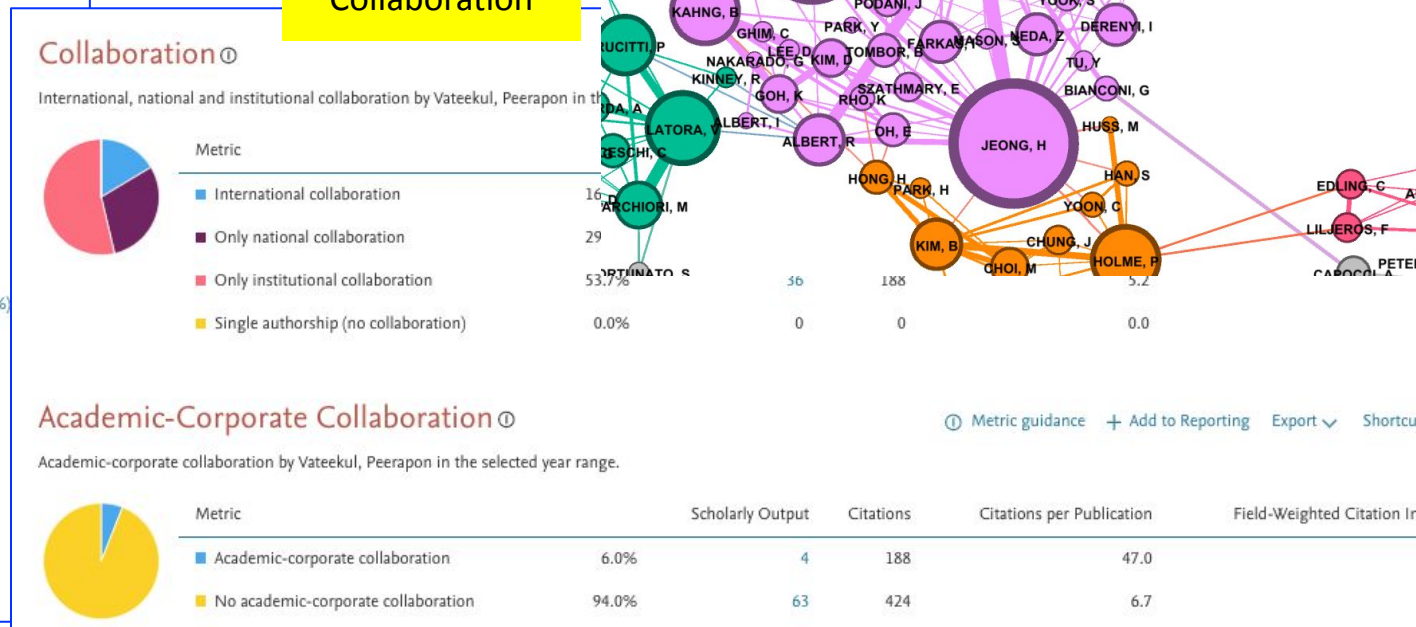
Example visualization / research question

- Co-authorship at the levels of individual, discipline, faculty, university, and country.
- Identify which academic fields are applying macl along with trends in their application.
- https://www.scival.com/home

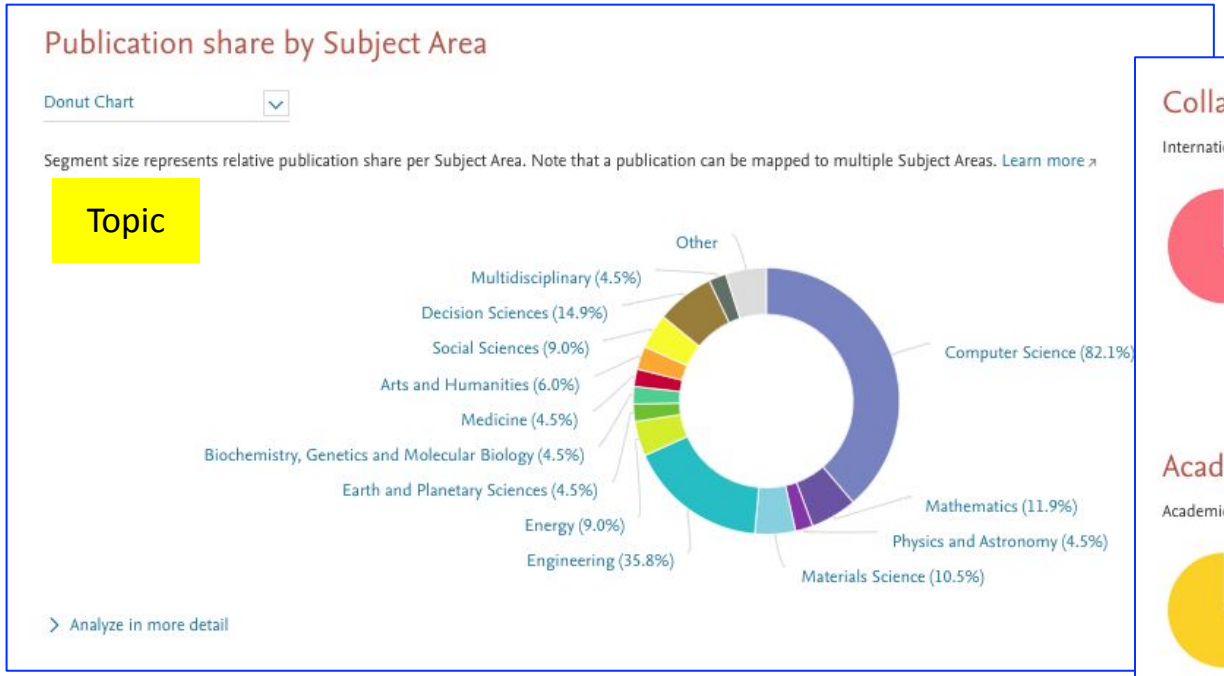
Co-author network graph



Collaboration



Topic



Scoring criteria (10%)

- The scoring criteria will be based on:
- **Completeness:** ensuring all 3 required components + data with web scraping
- **Project interestingness:** evaluated through:
  - Effort (e.g., additional data),
  - Creativity,
  - Execution,
  - Technical quality,
  - and other relevant factors.

Presentation & submission (5%)

- **Presentation Video**
- Upload the video to YouTube and set it to public. The video should be shared in the project channel on Discord by the deadline for other groups to view. The video length is 15 minutes, where the presentation quality will be a key factor in evaluation.
- The video should include:
  - 1. An explanation of the data used, including any additional data sources if applicable.
  - 2. A breakdown of each of the 3 required components, with a diagram illustrating how they interconnect.
  - 3. A demo showcasing interesting results from your data analysis.
- **Other Deliverables**
- Source code, presentation slides (in both PPT and PDF formats).
- Submission should include a link to a Google Drive folder shared on myCourseVille with “viewer” access for anyone.