### Problem statement

The objective of this project is to develop a machine learning prototype for clustering potential investments in bio-related companies based on the information extracted from a large set of unstructured PDF files. These files contain varying information focused on either scientific/biological aspects or financial/investment aspects of the companies. The challenges faced in this project include:

* The dataset comprises 600 unstructured PDF files with varying file sizes (ranging from 71KB to 50MB) and page counts (ranging from 1 to 200).
* Some of the files are corrupted or password-protected, which may hinder data extraction.
* The files contain a mix of finance company reports and general information on bio-related companies.
* No clustering variables have been specified, requiring the identification of appropriate variables to use in the clustering process.
* No investment criteria or priorities have been defined, necessitating the exploration of potential criteria to guide the clustering process.

The goal is to effectively clean and extract data from these files and determine the best clustering method to categorize the potential investments for further analysis and decision-making.

## The Approaches

The approach taken to address the problem was to first organize the dataset, and then perform clustering within each category. To achieve this, I cleaned the files and removed any corrupted PDFs. In order to determine which data should be extracted from the PDF files, I conducted a literature review to identify the most significant success-determining features for bio-related companies.

Subsequently, I experimented with various data reading and cleaning strategies, as well as different methods for extracting data from the PDF files. To assess the accuracy of my algorithm, I compared the data extracted using my algorithm with the data manually extracted from a randomly selected set of PDF files.

Other strategies could also be considered. For instance, if we were to define its values and priorities, it would be possible to perform clustering and provide different investment opportunity clusters, each with a unique ranking number.

### 

Given the time constraints, limited resources, and information availability for this project, I chose the approach I pursued based on the specific goals we aimed to achieve. Generally, I opted to organize the data before clustering, as the risk-reward profile for each category of companies could vary significantly. By organizing the data first, we ensured that the clustering process took into account the unique characteristics of each category, allowing for more accurate and meaningful results.

### A technical description of the prototype

The prototype consists of several steps, including file cleaning, reading PDF files, extracting data from the files, evaluating data extraction, and determining the optimal strategy and parameters for extracting each feature from the PDF files. Additionally, it involves organizing potential investments, experimenting with various clustering methods and hyperparameters, and evaluating the clustering using the silhouette score. Based on these results, the best clustering method and hyperparameters are selected, and the companies within each cluster are identified.

For a more comprehensive explanation, please refer to the detailed documentation provided within the code folder.

### Evaluation

I validated the data extraction by comparing it with a manually extracted dataset from 5% of the files, and I assessed the clustering method by calculating the silhouette score.

### Next

1\_Enhance Data Extraction and Evaluation Accuracy:

a. Explore additional tools for PDF parsing, AI-based question-answering systems, and data cleansing techniques.

b. Develop a larger, manually curated dataset to strengthen the data extraction evaluation process.

c. Identify and rectify common data extraction errors by learning from past mistakes.

d. Train a custom question-answering model tailored to our specific needs.

2\_Expand Data and Information Extraction:

a. Extract team-related data, such as team size, interdisciplinary composition, and connection to field leaders (Eros number).

b. Analyze the number of professional connections on LinkedIn and Twitter, paper citations, and networks of collaborating and investing companies.

c. Investigate relationships to trending topics on Twitter.

3\_Establish Company Priorities:

a. Collaborate with the team to identify and define the company's key priorities.

b. Evaluate factors such as market size, the level of innovation, and the company's focus on computational or non-computational approaches.

c. Explore potential collaboration opportunities within the industry.

4\_Experiment with Advanced Clustering Techniques:

a. Test various clustering algorithms and optimize hyperparameters to enhance the data analysis process.