**Education Challenge - Startup Genome**

For every technical module of the Master in Data and Decision Science of Sirius, a practical challenge needs to be developed by a group of students as part of their teaching process and part of their learning assessments. Therefore, the business question and the dataset provided by the partner company will define at what point in the teaching journey (sequence of technical modules below) the challenge will be worked, based on the technical and analytical skills required for the development of the work.

1. Business Intelligence and Statistical Analysis
2. Future Casting: Predicting the Future with Machine Learning and Statistics
3. Diagnostics and Optimization for Decision Science
4. Fundamentals of Data Engineering with Python and SQL
5. Design of Data Product & Service

Within this context, in order to submit the challenge, we request that you fill in the fields below for preliminary analysis by the Sirius teaching team.

**Briefly describe the business justification/motivation of the proposed challenge**

Present an overview of the context in which the project is inserted. Inform the problem (need) or opportunity that justifies why the project is carried out. It is recommended to describe the challenge / scalability / scope / rationale and benefits.

**Objectives**

The objective is to demonstrate the transformation that the project intends to achieve. The definition of objectives should be clear and will guide the actions of the data science project. In general terms, the business question and the central problem identifying the opportunity should be explored.

**Dataset**

Dataset link / Dataset Upload (preferably with a preliminary description of the data dictionary in order to facilitate the students' work). The analysis of the data by the students will allow evaluating the possible answers from the provided dataset.

---------------------------------------------------------------------------------------------------------------------------

Challenge from Startup Genome:

* Here we are sharing the challenge:
  + Patents data is very important when we are evaluating the startup ecosystem, as Patents help us to identify the scale of innovation happening in the ecosystem and it’s very interesting to look the data on sub-sector level.
  + Analysis on sub-sector level helps us to identify the new and emerging technologies in startup ecosystem.

We would like to divide the challenge (Analysis) in below categories

* Dataset
* [global patent database](https://drive.google.com/drive/folders/1vXzLp189I-BH27PsIHkg4LpRp3HpBTIa)
* [General Terms Link](https://docs.google.com/spreadsheets/d/1jrNn98RsUHDaUuRQZioN5NfJvlNPkzeW6QtEkUFQsV0/edit#gid=0)

**Challenge:**

* Number of Patents by Ecosystem by year
* Ecosystems that are moving fast
* Growth Rate
  + YoY
* Moving average
  + 3 Years
* Number of Patents by Ecosystem by Sub-sectors
  + Directions:
  + First step is to map patent [IPC](https://docs.google.com/spreadsheets/d/14_NqzhmB74M3GuiGDijg-83rLjqXhl4byYcx7RCwdvU/edit?usp=share_link) to sub-sectors
  + We have shared all patents [abstracts](https://drive.google.com/file/d/10k-uOlXjsoqBUWVloZ0O4cqEzSD3PP8f/view?usp=share_link) (Descriptions) which can be used to check potentially if any patent is falling in any subsector (subsector definitions provided)
  + Find a way to assign each patent of IPC to different subsectors based on some insights about a patent (Hint: Abstract can be used to apply machine learning, text preprocessing or any suitable method as per your understanding about the data)
  + When a subsector is identified for any IPC For eg. An abstract got classified under Life Sciences check that patent’s IPC and tag that patent’s IPC to Life Sciences.
* Rank each ecosystem with a set of criteria for
* “Overall” (sum of patent count of all sub-sectors )
* Downward citation score -
  + This can be achieved by checking similarity in patents abstract, generally certain counts of patent are recited with the same abstract.
* Patent Creation Strength of a Company
  + Check by company name ([List of Companies](https://drive.google.com/file/d/1u3zfGbTSLOX5auZ22iOA95gBxGGwouJT/view?usp=share_link) with name,city,country - provided) how many companies can match with applicants name ([Applicants Data](https://drive.google.com/file/d/1qB-EYRExL7WvPq_dLAbM_YVdQ4WYTtX-/view?usp=share_link) provided) in order to derive the count of patents produced by that company. (*this task may be more sensitive as names of Companies and Applicants may/may not have a Complete match*)
  + Sum the matched company names individual counts on ecosystem level to determine which ecosystems are on top.