**Pioneering Data Science Venture Capital Project**

*Data Science in Fintech = Data-Driven Decisions + Problem-Solving + Storytelling*

**Task #1: Building Profiles of Well Known VC Entities from Public Data**

# Project’s Overall Objective & Overview

* Create the Industry’s most encompassing, detailed, and continuously up-to-date **BERYL DBASE VC Investors’ Profiles** by identifying data-analytics challenges & the best tools to solve them.
* Identify the optimal methods to automate Text Mining process, NLP models, Machine Learning, Algorithms into a programmable, cohesive & scalable engine for both **BERYL VENTURES’** internal purpose and BERYL’s future **DBASE** **SaaS product**.
* The expected outcome of the Project includes a list of text mining algorithms per raw data source used for training models to extract the most descriptive dataset for Investor Profile.

PLEASE NOTE: We will provide the list of Investors that invest in VC to you. Please note that the list contains VC firms, family offices, hedge funds, foundations & endowments, pension funds, and other entities that invest in VC deals.

# Task #1 Objectives

* **Develop a thorough understanding of a well-known investor company (VC):**  Research their investment Thesis & Strategy, Sector & Industry focus, Partner’s background & pedigrees, number & Type of Companies they invested in, typical investment Ticket Size, City & State of their investments, notable activities & public speaking engagements.
* **Perform text mining process to an investor’s online platform(s):** Web scraping on social media such as LinkedIn and Twitter, their blog (if they have one), keeping up to date with any of their public speaking engagements, or access public data APIs who provide information on the investor’s activities.
* **Identify sources of raw data related to their investments**.
* **Collect and monitor the content they share:** Automatically collect any new articles, photos, or YouTube videos they post, any comments or likes they make on target websites.
* **Correlate and cluster their engagements** with other professional peers and/or companies.
* **Identify potential areas of interest or collaboration** based on the data collected.

**\*\* IMPORTANT**:

PLEASE FORMULATE ANY QUESTION(S) YOU MAY HAVE AND/OR BRING FORWARD ANY IDEA / FEEDBACK REGARDING THIS TASK. YOUR FEEDBACK IS VERY IMPORTANT TO US IN ORDER TO FINE TUNE & OPTIMIZE THE PROCESS.

# Activities but not Limited

1. **Collect and prepare the data:**

a. Identify the sources of data that are available for these well-known VCs [Task 1 - Sample Set - Well-known VC/PE](https://docs.google.com/spreadsheets/d/1N-isW5E9DO8NzU9tLCfhBC4yomGzWXJa0sWyi4euzMM/edit?usp=sharing)(Vidak provided a list of public websites from his experience for you to select from [Target Websites](https://docs.google.com/spreadsheets/d/1FfT1rO3kerS28KtdPw-3bFfHuaotybDnRTL_WXzP6jk/edit?usp=sharing) but you conduct further research, including public SEC filings.)

b. Acquire and integrate the data into a single usable format.

c. Evaluate the quality of the data, clean them and modify as necessary.

d. Transform the data into a format that is suitable for data mining.

1. **Choose the data mining techniques and tools:**

a. Identify the data mining techniques, ML Libraries that are the most suitable for this project and you are comfortable to use.

b. Choose the appropriate data mining tools and software you are comfortable to use.

c. Decide on the parameters and settings for the data mining process.

1. **Perform the data mining:**

a. Run the data mining process and analyze the results.

b. Validate the results to ensure accuracy and reliability.

c. Make modifications and adjustments to the data mining process as needed.

1. **Present the results:**

a. Summarizing the data mining results in an understandable, easy-to-read format.

b. Communicate findings to Beryl senior management, brainstorm the next step.

page break