# **Vision Statement – R&D**

Eviatar Nachshoni 316160464

Yosef Danan 211838420

Nir Son 323918599

**Goals:**

The problem that we want to solve in our project is to identify the author of a book given cluster of books with the same genre with different authors.

First, we define the scope of our problem, we decided to focus on recognize author by his writing, style writing, for this we need:

* We already arrange the data, such that for every author have thousands of paragraphs, for that we divide every book to chunks of 500 words.
* We convert the sentences to features based on chunks, n-grams, POS, sentence average, average of punctation per sentence.
* Overview some articles for get information on network's architecture and features for style identification, we will try to implement the architecture and merge between different suggestion of architecture for to achieve better result than the existing.
* If the time allow us, we will also try to solve the problem by using unsupervised learning techniques.
* We will evaluate our model by precision, accuracy, recall, and latency of model on one prediction, we will try to check our model on different size of authors and try to evaluate our model by this data set.

**Project scope:**

We will focus on identifying author only by his style writings, else our model will focus only on books in English languages, we won't predict author's location or period, or topic/genre of the text and we won't try to inspect our model on some languages.

**Data set**:

we found data set of books in English that have tens of thousands books in English from different authors (from different demographic places Russian, USA and etc. and different topics), we upload the data after we download and clean her to [here](https://drive.google.com/file/d/1UnTLPc0pnxDZUso-ruCu_egOnHHkJ0sh/view?usp=sharing), the data contains almost 8,000 authors, but for some authors have only few books, then we remove all the authors that have less than 5 books, hence we left with 300 authors.

We already built some feature that already proved that represent style writing, features like sentences length, average sum of stop- word use in sentence, etc.

We work on Jupyter Notebook, and we put all our notebooks in [GitHub](https://github.com/YD5463/Author-Identification).

Also, we have one more [dataset](https://pan.webis.de/data.html#c50-attribution) that we will use to see how our models perform when adding more labels after retraining.

**Our work plan:**

we will start with the problem with two different author and the same main genre, that way, it is attempted to minimize the topic factor in distinguishing among the texts.   
At first, we will use simple features like word frequency and simple ML models

after that we will increase the count of the authors but with attention to minimize the topic factor and use more advanced deep learning models.