**The Motivation Behind our Project**

Our digital world enables fast circulation of information in all aspects of human life. Patents are seen as sentinels of *valuable information* in a sense that a granted patent has been going through an approval process and hence may represent ideas in a correct, unique and non-reproduced way. Additionally, a patent allows the retrieval and organization of such information in a precise manner, as these documented grants of intellectual property follow means of standardization and categorization. Patents in general are described to provide two important functions, *protection* and *disclosure*.

As the World Intellectual Property Organization (*WIPO*) emphasizes, patents may be granted on territorial principles, but the information disclosure is available on a global scale. This bears both problems, but also opportunities: being able to investigate the entire global body of innovative output may yield data covering many real-world business and research problems. To name just a few suggestions, one could for example think about investigating whether the innovative output of a particular industry is shrinking. Historical research on the impact of events such as turmoil, economic crises or conflicts on a territory’s innovation output would be another example. As a last suggestion, any body of education may be interested which current trends are emerging – a certain niche market may soon be asking for highly qualified and specialized talent, and as such may not be able to provide this degree of education by itself during its growth.

However, as there is a plethora of patents available and the proper investigation of these may consume a large amount of human resources spent on a rather simple, reproduceable assignment, we investigate patents with methods of computational extraction and analysis. First and foremost, our initial questions of interest have been identified as:

***“Are the patents that are being filed in the Special Administrative Region of Hong Kong (HKSAR) following underlying patterns that go beyond a usual classification of industries and such features on the surface? How can a niche market with high potential be described, based on this assumption?”***

**Project Scope and Limitations**

Since this endeavor serves as a final project over a full-time machine learning course, we aimed to accompany our relevant data (e.g. patents) throughout the entire value creation chain that is inherent in data science and machine learning: finding, extracting, exploring, and preparing the materials before training and applying models on them was crucial for us to understand possibilities, necessities and shortcomings in a thorough manner.

The limitations of this project were defined based on a rather short time frame of execution, and we preferred having minimum viable, and presentable, results over a holistic body of data. The method allows an easy widening of the scope, and by relying on Python with well-documented approaches to data analysis and natural language processing we believe that the reproducibility of our project is granted, and thus shall result in congruent findings. Hence, widening the territories under consideration as well as the time frame is highly recommended as a next step. For the scope of this project, we started our investigation with *all patents that have been filed under WIPO as a governing body in the HKSAR in the past twelve years.* By choosing this territory, we are certain to contribute to a point of interest for the society that we live and work in, instead of looking at other territories. The timeframe of the past twelve years was chosen by a trade-off between a lean, easy to handle dataset and the usability of such. Resulting in a set of full data comprising over fifty thousand fully-documented patents, we believe to have reached a critical mass to perform actions such as document classification and quantitative analyses, while the feasibility of the project given a tight deadline is granted.

**Methodology**

*Data extraction*

*Preparation*

*Explorative Data Analysis*

*Natural Language Processing: creating a classifier*

*Natural Language Processing: applying a classifier*

**Outcome**

**Next Steps**