DETERMINE BUSINESS OBJECTIVES

- ## BACKGROUND

  Obtener la información sobre la situación de negocio (trabajaron en esto en sesiones pasadas)

  - **SF works at *Instituto Data Science*, an institute created by *Universidad del Desarrollo* and *Centro I+D de Telefónica*.**

  - **SF is located in Chile, which local time is +2 hours ahead of Mexico's local time**

  - **Movistar is the company that provides the datasets and it has 30% of the market share in the telecommunications industry of Chile**

  ¿Cuál es la información del socio formador que es relevante para el contexto del proyecto que todos los interesados deben saber?

  They work at the Institute for Data Science in Chile. They have data from the Movistar Company.

  The project is focused in the Metropolitan area of Santiago, this will allow a better delimitation of the problem and more consistent proposals about the transport system.

- ## BUSINESS OBJETIVES

    - *Generate MODs (Matrices Origen Destino) for the region of Santiago Chile based on mobile data captured by Movistar so that an analysis and a set of insights can be reported*

    - *GitHub repository hosting all of the code that the model uses to generate the desired MODs**

    - *Generate a prediction using deep learning techniques*

    - *SF wants to understand the behavior of the trips inside of a particular region (metropolitan region)*

    - *Some possible questions on the business are:*

        - *1 What are the hours where people travel the most?*
        - *2 When are the holidays in this country?*
        - *3 How are the antena distributed ?*

- ## BUSINESS SUCCESS CRITERIA

    - *The socio formador could determine if after the project, we are getting enough information to make improvements in the region's mobility, such as changes on the public transport.*

ASSETS SITUATION

- ## INVENTORY OF RESOURCES

1.- *EXPERTS:*

Dr. Loreto Bravo (Director of Data Science Institute UDD/100 and our SF):

- ☐ Technical knowledge on phone data and mobility.
- ☐ Information about some external aspects of Chile.
- ☐ Tips for using data to make decisions and value information.

Dr. Benjamín Valdés Aguirre (Reto Coordinator 1 and Module 2 professor):

- ☐ Technical knowledge from data analysis and Machine Learning.
- ☐ Tips for ETL and EDA process.
- ☐ Best libraries for python coding.

MTI Eduardo Daniel Juárez Pineda (Reto Coordinator 2):

- ☐ Technical knowledge about CRISP-DM
- ☐ Adaptation of CRISP-DM to a project management methodology
- ☐ Tips for best administration project phases.

Dr. Ismael Solís Moreno (Module 1 and Module 4 professor):

- ☐ Technical knowledge of Big Data.
- ☐ Best solutions and methods with Spark.
- ☐ Technical knowledge of Cloud Computing.
- ☐ Knowledge of Tableau and AWS services.

Dr. José Antonio Cantoral Ceballos (Module 3 professor):

- ☐ Technical knowledge on natural language processing
- ☐ API management.
- ☐ Methods for transcriptions, use of chatbots and texts analysis.

Dr. Carlos Alberto Dorantes Dosamantes (Module 5 professor):

- ☐ Technical knowledge on statistical hypothesis testing.
- ☐ Statistical methods for best interpretation with data analysis.
- ☐ Technical knowledge on data analytics.

2.- DATA:

Datasets given by the socio formador

- ☐ *We have 2 datasets with information about codified (secured) data of cell phones for mobility of the population in Santiago de Chile, Chile.*

Public datasets

- ☐ *We could investigate in public pages for additional information datasets in case of requirement for other analysis, conclusions and perspectives.*

Some public and real information available on the Internet.

- ☐ *On the internet, we could search for specific data of some cases or aspects from the location in Santiago de Chile.*

3.- COMPUTING RESOURCES (HARDWARE):

1. Teammate's laptops
2. $400 USD credit of AWS computing resources.

4.- SOFTWARE:

1. Google Collaboratory
2. Google Online tools (Docs, Sheets)
3. Github
4. Spark
5. PySpark
6. AWS tools
7. Slack
8. Tableau
9. Python
    a. Pandas
    b. Numpy
    c. Tensor Flow
    d. MatplotLib
10. Markdown tools

- ## REQUIREMENTS, ASSUMPTIONS AND CONSTRAINTS

    - *We have the necessary permissions to use this data. All the phone numbers are anonymized so nobody can know who the data belongs to.*

    - *We assume that the data that we have is truthful and enough to develop a solution to the challenge.*

    - *We are limited to using some equipment to develop the project and we could use AWS, Google collaboratory and other tools that may help us.*

    - *If we need additional resources of information the only way to get is from Internet, using public data sets or another data from official sources and organizations with a good private policy about the use of the data.*

    -

- ## RISKS AND CONTINGENCIES

    - ***Language communication barriers between team members***
        - *Standardization of the English language for any document or communication made by the team*
    - ***Delayed feedback from SF***
        - *Contact SF prior to the request of feedback to avoid any delay in responses*

- ## TERMINOLOGY

    - *SF (Socio Formador): Entity that proposes the challenge for the course and provides guidance for its solution.*

    - *PySpark: interface for Apache Spark in Python.*

    - *Python: Coding language.*

    - *AWS: A set of tools for different technologic areas, each service has a lot of functionalities and uses in the industry and investigation.*

    - *Matplotlib: library used to plot graphics***.**

- ## CRISP-DM ADAPTATION

    - *The deliverable for COST AND BENEFITS has been removed due to the academic nature of the project.*

DETERMINE DATA MINING GOALS

- ***DATA MINING GOALS***
    - Transformar los objetivos de negocio en términos técnicos**.**
    - *At the end, we must obtain the origen destino matrix (MOD), so we can generate knowledge that is useful for the socio formador.*
- ***DATA MINING SUCCESS CRITERIA***
    - *We can make comparisons between the matrix generated by the Socio Formador and the matrix generated by us. This way we will be able to measure the success of our analisis.*

PRODUCE PROJECT PLAN

- **PROJECT PLAN**
  - **Descargar los datos**
  - **Leer el dataset y entenderlo**
  - **Junta intermedia para checar avances y actualizar documentos**
  - **Explorar el dataset (EDA)**
  - **Generar gráficas(EDA)**
  - **Deshacernos de valores outliers y NaN**
  - **Junta intermedia para checar avances y actualizar documentos**
  - **Generar matrices**
  - **Generar análisis de la matriz MOD**
  - **Junta intermedia para checar avances y actualizar documentos**
  - **Considerar propuesta para desarrollar una solución de utilidad para el socio**
  - **Escoger una propuesta para desarrollar una solución**
  - **Junta intermedia para checar avances y actualizar documentos**
  - **Generar un modelo**
  - **Evaluar el modelo**
  - **En caso necesario, generar nuevo modelo**
  - **Junta intermedia para checar avances y actualizar documentos**

- **INITIAL ASSESSMENT OF TOOLS, AND TECHNIQUES**
  - Debe incluir la selección inicial de las herramientas de minería de datos
    - **spark**
    - **python**
    - **Pandas**
    - **numpy**
    - **matplotlib**