Group Proposal

What problem did you select and why did you select it?

To predict which previously purchased product would be in a Instacart customer next order.

This problem is very practical and interesting. Working on this problem, we can use transactional data to develop models that predict which products a user will buy again, try for the first time, or add to their cart next during a session.

What database/dataset will you use? Does it need to be cleaned?

The Instacart Online Grocery Shopping Dataset:

- The dataset contains 3 million orders from 200,000 users and for each user, it provides between 4 and 100 of their orders.
- All data are obtained from Kaggle Competition website and a relational set of .csv files which describe customer orders over relative times. Each entity (customer, order, department, etc.) has a unique id.
- A blog post has additional information about this dataset.

The dataset needs to be cleaned, because the subsets containing various of information needs to be merged and there exists missing value to be process and we need to perform feature selection.

What data mining algorithm will you use? Will it be a standard form, or will you have to customize it?

- First, we need to use several packages like Numpy, Pandas, Scipy to preprocess the data sets and to do EDA. Then, we will use some machine learning algorithms from sklearn package, such as decision tree and random forest, to train the data and to build models.
- We will use the standard form algorithm.

What software will you use to implement the network? Why?

Use Python (Pycharm) for coding, SQL(MySQLWorkbench) for database querying and Git (Github) for collaborating. Since we have to merge several data sets using SQL first, then we need to clean and preprocess the data and finally do the analysis and model using Python.

What reference materials will you use to obtain sufficient background on applying the chosen network to the specific problem that you selected?

Since the project we choose is from Kaggle, we would refer to the description in Kaggle.
Citation:
"The Instacart Online Grocery Shopping Dataset 2017"
Accessed from https://www.instacart.com/datasets/grocery-shopping-2017 on <date>

How will you judge the performance of your results? What metrics will you use?

We will use mean F1 score, ROC curve, Gini coefficient, accuracy score and confusion matrix for evaluation.

Provide a rough schedule for completing the project.

10/23 - 10/30: problem specification and understanding
10/30 - 11/13: data preprocessing and EDA
11/13 - 11/27: model development/data mining and evaluation
11/27 - 12/04: project presentation