MLDM Project Proposal

# Student

Egor Dudyrev (individual project)

# Dataset

M5 Forecasting: <https://www.kaggle.com/c/m5-forecasting-accuracy/data>

About dataset:

The dataset consists of Walmart sales information: when, what and at what price an item was sold. The task is to predict amount of specific items sold in specific store for each of next 28 days.

The data covers stores in three US States (California, Texas, and Wisconsin) and includes item level, department, product categories, and store details. In addition, it has explanatory variables such as price, promotions, day of the week, and special events.

# Research goals

1. Compare different model predictions (e.g. Linear Regression, SVM, Random Forest, Gradient Boosting algorithms)
2. Check how sales correlate with special events (null hypothesis here is that most sales are caused only by price, promotions and weekends and not the other features)
3. Construct an items ontology based on item\_id, dept\_id, cat\_id. Try to use this information with FCA prediction and interpretation.

# Roadmap

1. Create baseline model based on amount of previously sold items
2. Create plain dataset from all the files and create more difficult model on it (e.g. CatBoost with selected categorial features)
3. Check Kaggle discussions for the additional datasets (e.g. weather information)
4. Create an ontology of items and estimate the sales with FCA
5. Create ensemble model based on all previously used models
6. Ensure that current set of selected features gives the best performance
7. Interpret the model (find the most important features and some insights on the data)
8. Write the report