# Given Data

The following data of a delivery service for groceries is given:

1. **orders**:   
   Contains one line per order with the following columns:

user\_id: Unique ID for each customer  
order\_id: Unique ID for each customer

order\_date: Timestamp for each order (date and time of day)

1. **order\_products:**One line per order item, i.e. per order there are as many lines as items (=articles) in the order  
   Columns:

order\_id:

product\_id:

1. **products, departments, aisles:**Maps product\_ids to product names. Additionally, products are hierarchically organized in aisles and departments. The mappings of department\_ids and aisle\_ids to names is defined in the table with the corresponding name.
2. **tips25\_trainingdata**:   
   One line per order. Contains the following columns:

order\_id:   
tip: Was a tip given tot he delivery driver? (yes/no)

The last order of each user is missing in this table. These orders form the test set for which you don’t know the tipping behavior.

# Goal

The general goal ist o develop a prediction model which predicts for an incoming order, whether or not a tip will be given upon delivery. The prediction for a specific order can potentially make use of all information available at this point in time, i.e.

* All information (including tipping behavior, articles contained in the order, …) of all orders oft he same user or other users that were placed earlier.
* All information except the tipping behavior about the order for which the prediction ist o be made, i.e. timestamp, articles contained, …

You will be guided towards this goal by subtasks which will be published over the course oft he semester.

# Subtasks

1. Familiarize yourself with the data set, do some exploratory data analysis using „Project\_start.jpynb“ as a starting point.
   1. If you run into performance problems you can limit your analysis to a subset of users and only their orders and order\_positions.
   2. Create a diagram or table showing the percentage of orders with tips over time. (You can orient yourself at the diagram with „number of orders“ over time in the „Project\_start.jpynb“. )
2. Compute what proportion of predictions will be correct (=‘accuracy‘) on the training set if one uses the following simple prediction strategies:
   1. Always predict the majority class. (Trivial model with 0 input variables.)
   2. Always predict that the tipping behavior will be the same as in the previous order of the same user. (For the first order of a user use (i))
   3. Create a decision tree model (maxdepth=3) using only the time of day and the day of week as inputs.