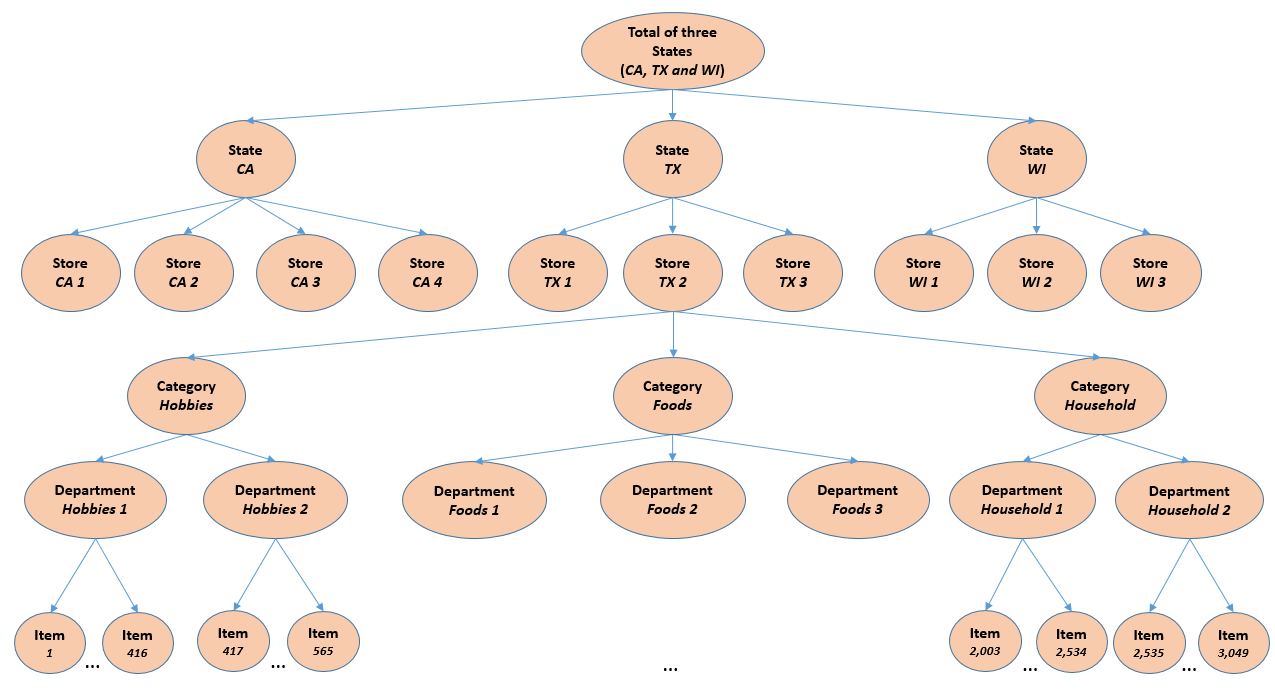
# **Data Description**

The M5 dataset, generously made available by **Walmart**, involves the unit sales of various products sold in the USA, organized in the form of **grouped time series**. More specifically, the dataset involves the unit sales of **139 products (instead of 3,049 in the original M5 competition)**, classified in **3 product categories** (Hobbies, Foods, and Household) and **7 product departments**, in which the above-mentioned categories are disaggregated.  The products are sold across **ten stores**, located in **three States** (CA, TX, and WI).



**Figure 1: An overview of how the M5 series are organized.**

The historical data range from **2011-01-29** to **2016-05-22**. Thus, the products have a (maximum) selling history of 1,913 days (**test data of h=28 days not included**).

The M5 dataset consists of the following **three (3) files**:

**File 1: “*calendar.csv”***

Contains information about the dates the products are sold.

* *date*: The date in a “y-m-d” format.
* *wm\_yr\_wk*: The id of the week the date belongs to.
* *weekday*: The type of the day (Saturday, Sunday, …, Friday).
* *wday*: The id of the weekday, starting from Saturday.
* *month*: The month of the date.
* *year*: The year of the date.
* *event\_name\_1*: If the date includes an event, the name of this event.
* *event\_type\_1*: If the date includes an event, the type of this event.
* *event\_name\_2*: If the date includes a second event, the name of this event.
* *event\_type\_2*: If the date includes a second event, the type of this event.
* *snap\_CA*, *snap\_TX*, and *snap\_WI*: A binary variable (0 or 1) indicating whether the stores of CA, TX or WI allow SNAP[[1]](#footnote-1) purchases on the examined date. 1 indicates that SNAP purchases are allowed.

**File 2: *“sell\_prices.csv”***

Contains information about the price of the products sold per store and date.

* *store\_id*: The id of the store where the product is sold.
* *item\_id*: The id of the product.
* *wm\_yr\_wk*: The id of the week.
* *sell\_price*: The price of the product for the given week/store. The price is provided per week (average across seven days). If not available, this means that the product was not sold during the examined week. Note that although prices are constant at weekly basis, they may change through time (both training and test set).

**File 3: “*sales\_train.csv*”**

Contains the historical daily unit sales data per product and store.

* *item\_id*: The id of the product.
* *dept\_id*: The id of the department the product belongs to.
* *cat\_id*: The id of the category the product belongs to.
* *store\_id*: The id of the store where the product is sold.
* *state\_id*: The State where the store is located.
* *d\_1, d\_2, …, d\_i, … d\_1913*: The number of units sold at day *i*, starting from 2011-01-29.

# **Model Description**

The following table lists the different levels at which model training could be performed. For example, at model level 10, it could be assumed that each item has different sales patterns but those patterns are similar across different stores, so models could be built per item. Since there are 139 items being sold across 10 states, there will be 139 models (1 for each item) and each model will be trained on 10 timeseries corresponding to the sale of the item in the 10 stores. At other model levels, there could be varying number of time series for training each model depending on the breakdown of the dataset as outlined in Figure 1 above.

**Table 1: Number of models per level.**

|  |  |  |  |
| --- | --- | --- | --- |
| **Level**  **id** | **Model Level** | **Number of models** | **Number of raw time series for training each model at this level** |
| **1** | Predict unit sales of all products, for all stores/states | 1 | 1,390 |
| **2** | Predict unit sales of all products, for each State | 3 | variable |
| **3** | Predict unit sales of all products, for each store | 10 | 139 |
| **4** | Predict unit sales of all products, for each category | 3 | variable |
| **5** | Predict unit sales of all products, for each department | 7 | variable |
| **6** | Predict unit sales of all products, for each State and category | 9 | variable |
| **7** | Predict unit sales of all products, for each State and department | 21 | variable |
| **8** | Predict unit sales of all products, for each store and category | 30 | variable |
| **9** | Predict unit sales of all products, for each store and department | 70 | variable |
| **10** | Predict unit sales of product *x*, for all stores/states | 139 | 10 |
| **11** | Predict unit sales of product *x*, for each State | 417 | variable |
| **12** | Predict unit sales of product *x*, for each store | 1,390 | 1 |

1. The United States federal government provides a nutrition assistance benefit called the Supplement Nutrition Assistance Program (SNAP). SNAP provides low income families and individuals with an Electronic Benefits Transfer debit card to purchase food products. In many states, the monetary benefits are dispersed to people across 10 days of the month and on each of these days 1/10 of the people will receive the benefit on their card. More information about the SNAP program can be found here: <https://www.fns.usda.gov/snap/supplemental-nutrition-assistance-program> [↑](#footnote-ref-1)