Chasing Instacart

Predicting who in the checkout line are Instacart shoppers

Shelly Hsu, Noah Randolph, Asha Anju

Data Sources

dunnhumby

instacart

- **Dunnhumby: The Complete Journey**
  - https://www.dunnhumby.com/sourcefiles
  - All of a household's purchases within actual grocery stores
- **Instacart**
  - https://www.kaggle.com/c/instacart-market-basket-analysis/data
  - The dataset is anonymized and contains a sample of over 3 million grocery orders from more than 200,000 Instacart users.

Variety and Volume

- Instacart dataset
  - **33 million** product purchases
  - **3 million** orders
  - **200,000** Instacart users

dunnhumby

- Dunnhumby dataset
  - **2.5 million** product purchases
  - household level transactions over two years from **2,500** households

Goal

- Predict which shoppers are from Instacart (or similar services) to help grocers compete with their own services
  - Competition should lead to cheaper deliveries for the consumer

amazon fresh

Walmart Pickup · Grocery

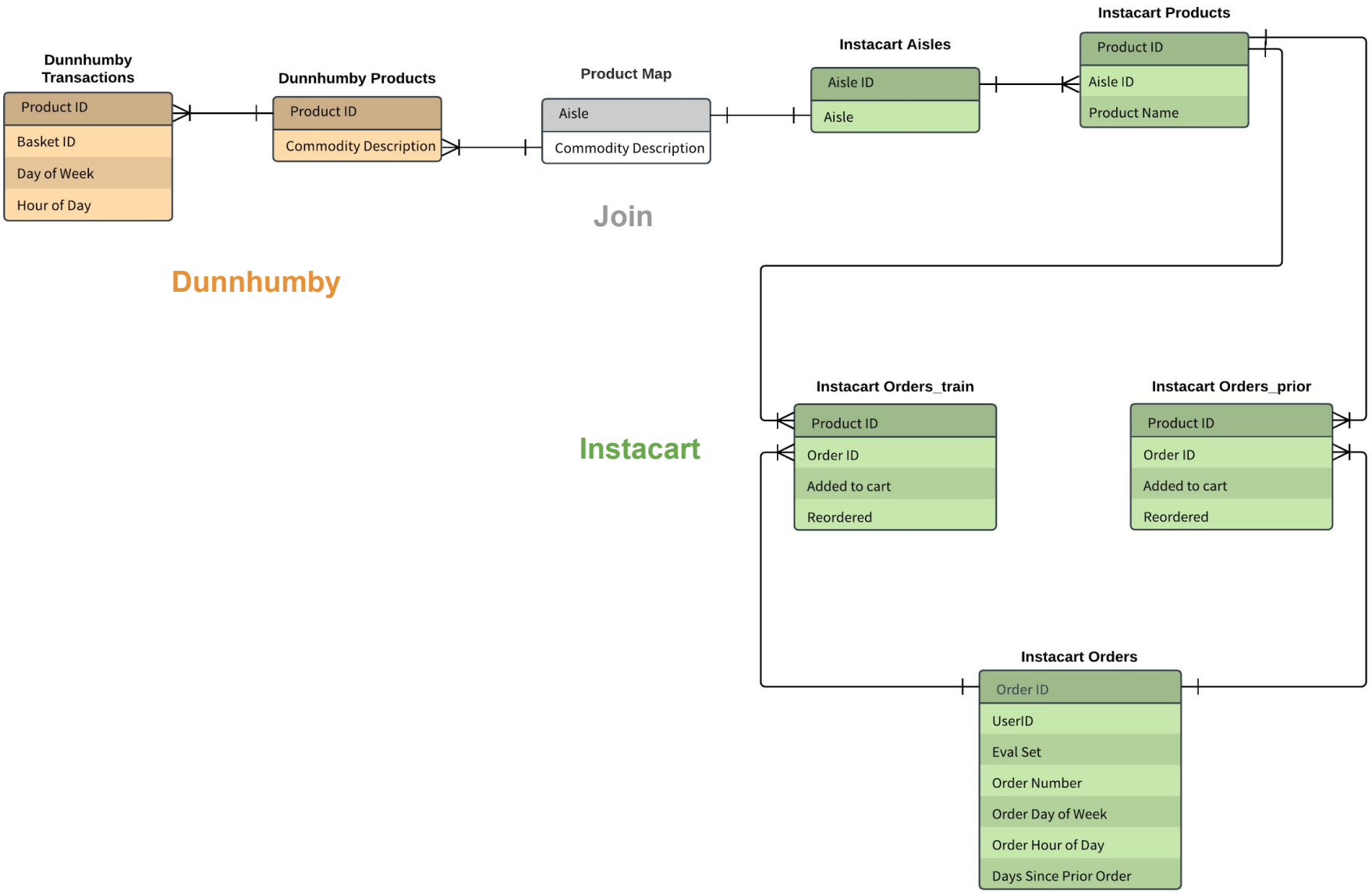
Architecture

- **End-to-end execution**
  - Bash script of nested scripts bash and SQL scripts
    - **Data ingest**
      - Nested bash script loads a data lake of 17 data files from S3 to HDFS, stripping headers (downloaded from S3 to ensure data remains available)
      - 10 Hive SQL tables created from HDFS data lake
    - **Data processing**
      - Instacart and Dunnhumby Separate
        - 22 Hive SQL tables created in Parquet format
        - joins as shown in ERD (upcoming slide)
        - convert product column to 1-hot encoded format for machine learning
        - format and join day of week and hour of day into 1-hot table
      - Union Instacart and Dunnhumby tables for one export to Python
    - **Machine Learning** (Continued on next slide...)

Architecture (continued)

- **End-to-end execution** (...continued from previous slide)
  - Bash script of nested scripts bash and SQL scripts (...)
    - **Machine learning**
      - Import table sections into data frames with PyHive
      - Rejoin into array, shuffle and split data into training and test arrays
      - Train model as a Multinomial Naive Bayes classifier using scikit-learn
      - Predict and score test data
      - Save model, test data, and test labels in Pickle
    - **Storing and reporting results**
      - Store test predictions, labels, and data as a data frame and write to .csv
      - Store in HDFS
      - Create Hive table of results data
      - Display results with Tableau
    - **Auxiliary process for EDA with Tableau**
      - Create auxiliary table for Tableau visualizations
        - 1-hot encoding not as good for Tableau as standard column format

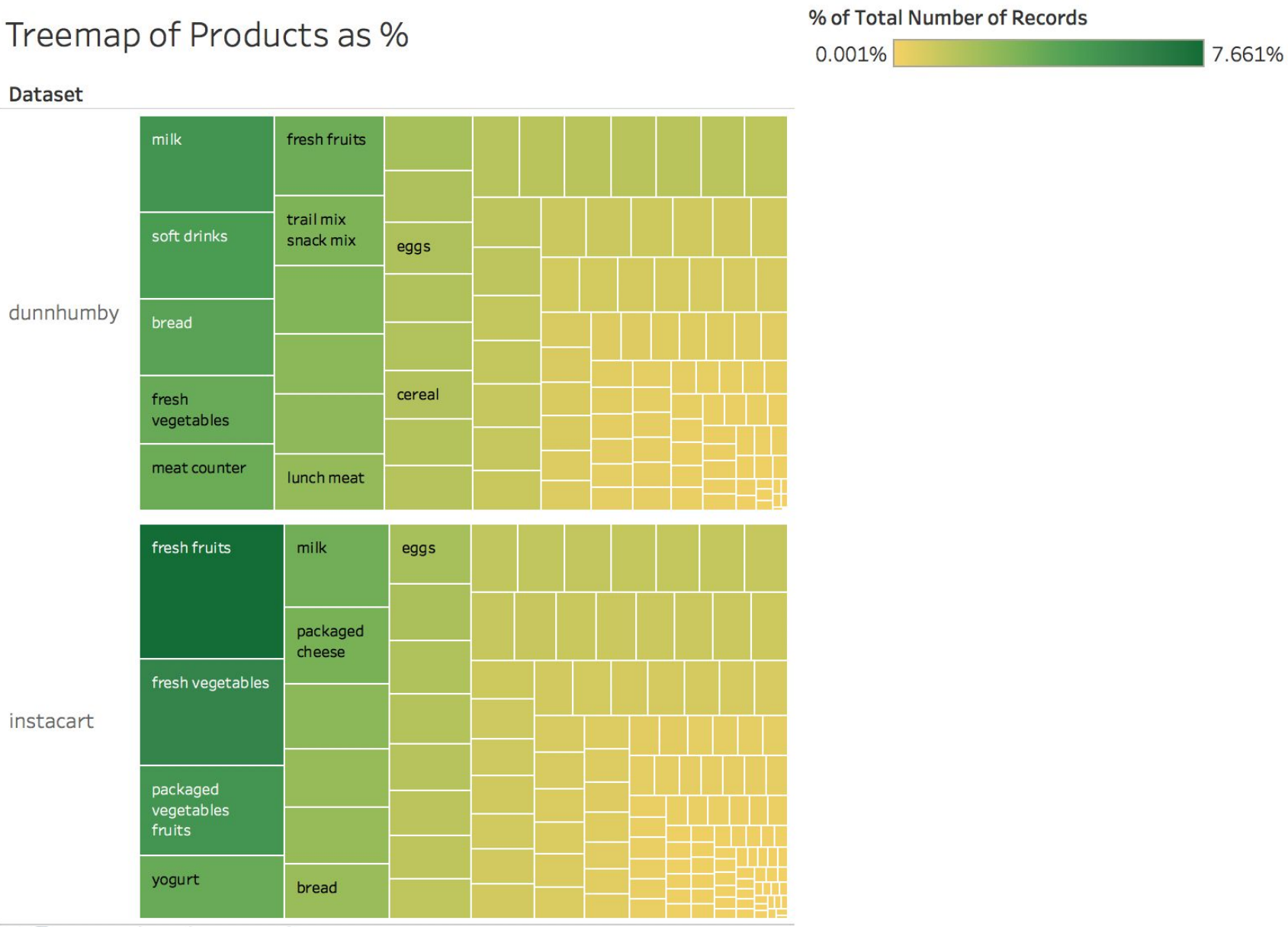
Entity Relationship Diagram Building up to A Dataframe



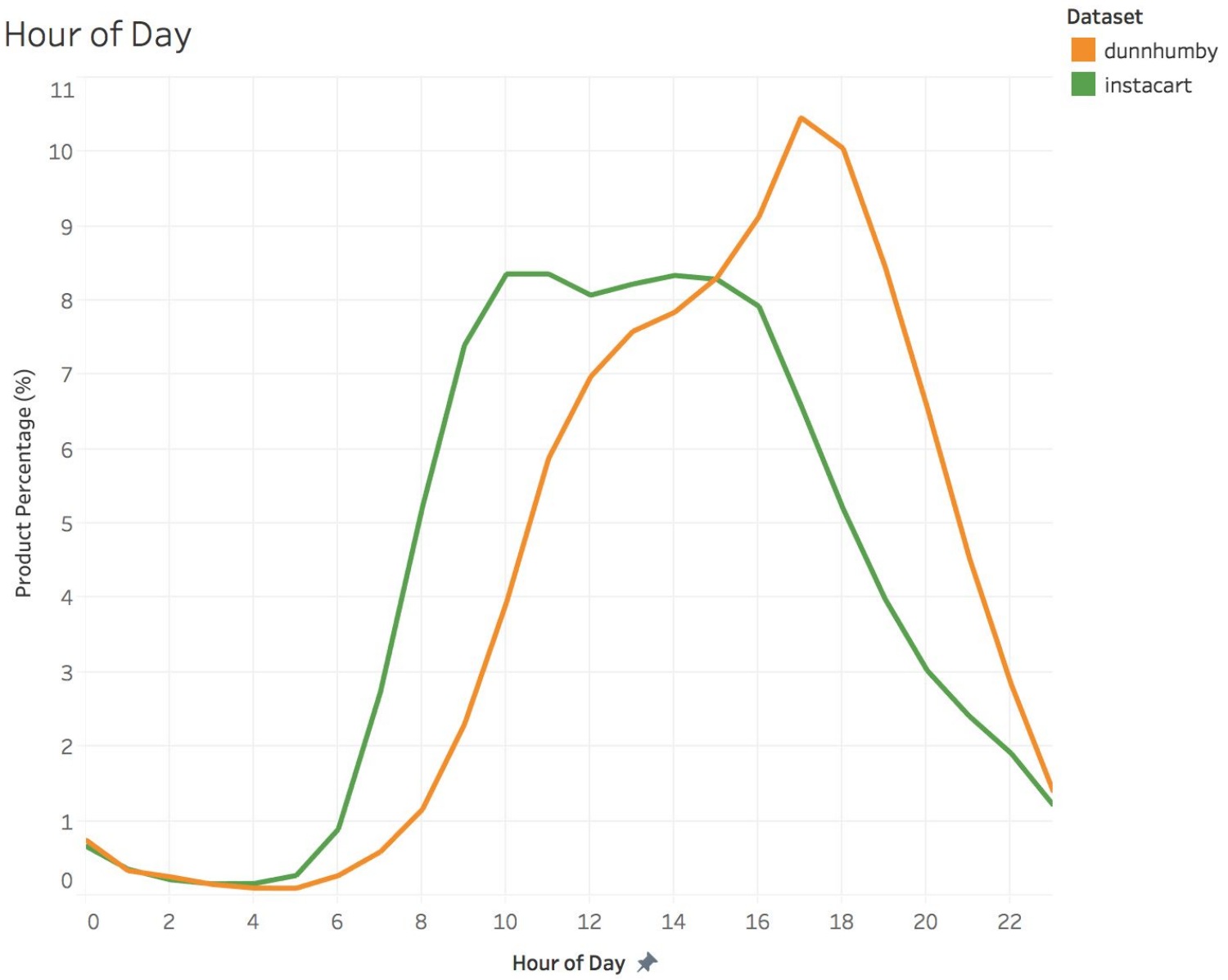
Final Table Orders for Analysis

Variable Name	Basket_id	Day_of_ week	Hour_of_ day	Product Aisles (n=106)				Dataset
Example Values	100 1000 269849 etc.	0 1 2 3 4 5 6	0 1 2 … 22 23	Air_ fresheners_ candles	candy_chocolate	...	yogurt	Instacart dunnhumby
				Null, 1, 2, 3, etc...				

Exploratory Data Analysis

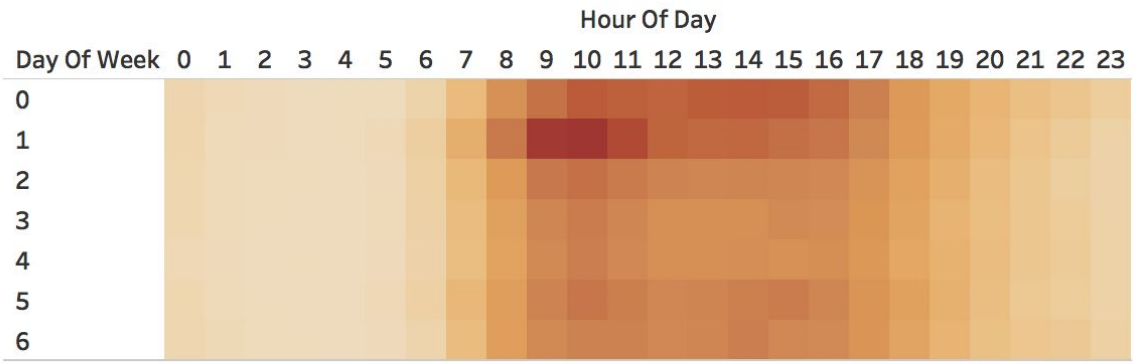


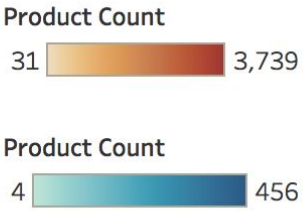
Exploratory Data Analysis



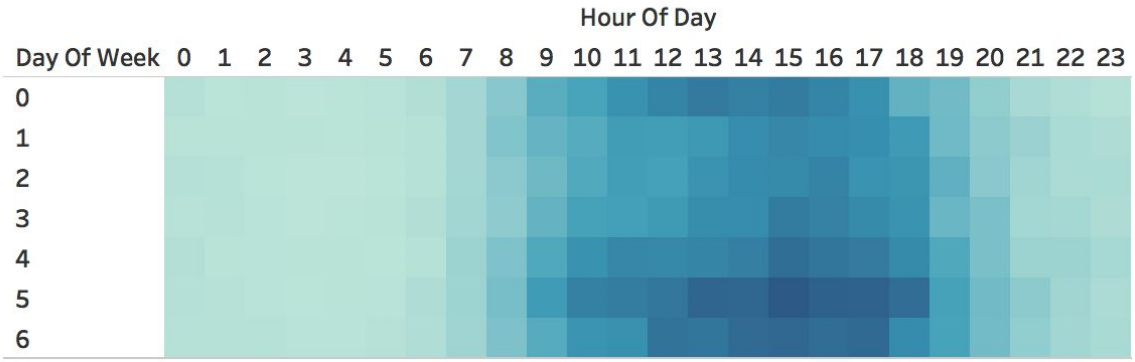
Instacart Days of the Week

Instacart Coffee



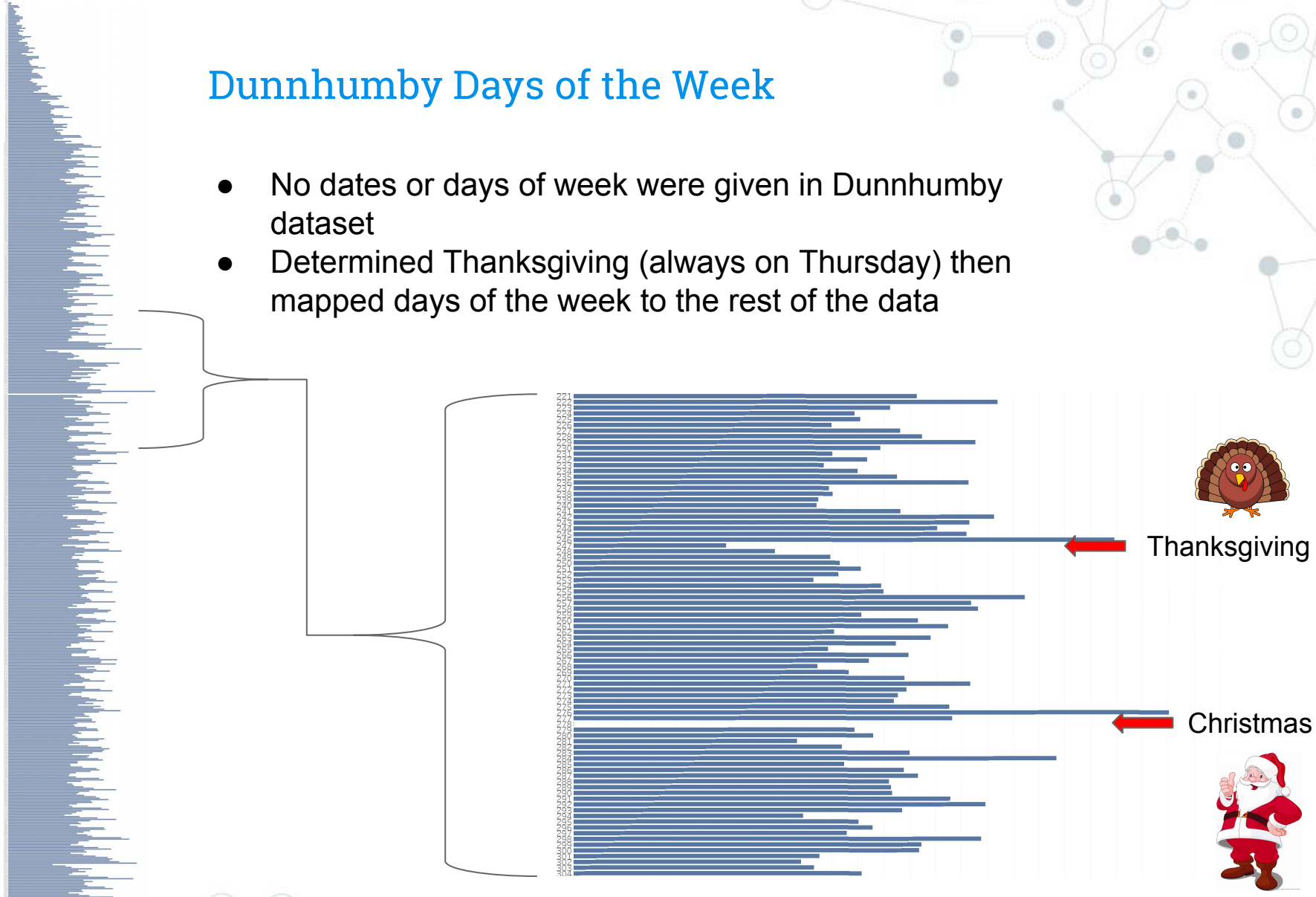


Instacart Spirits



Dunnhumby Days of the Week

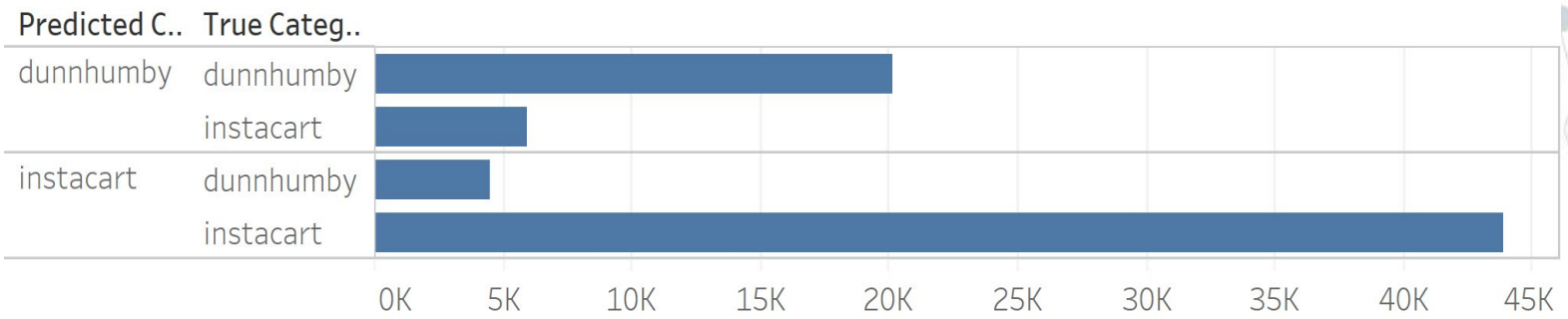
- No dates or days of week were given in Dunnhumby dataset
- Determined Thanksgiving (always on Thursday) then mapped days of the week to the rest of the data

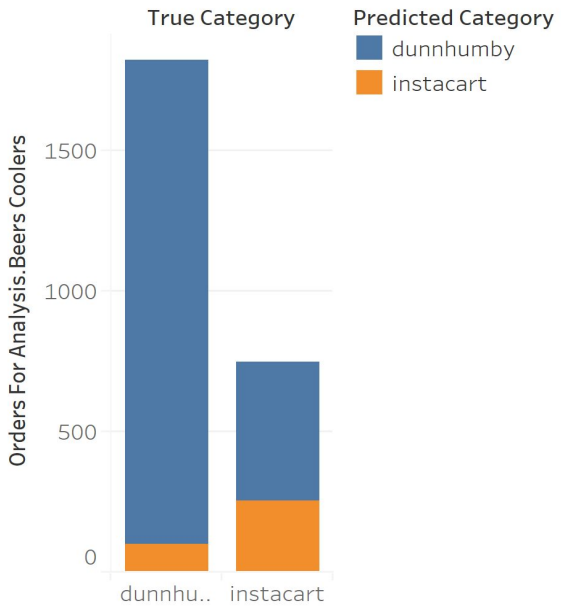


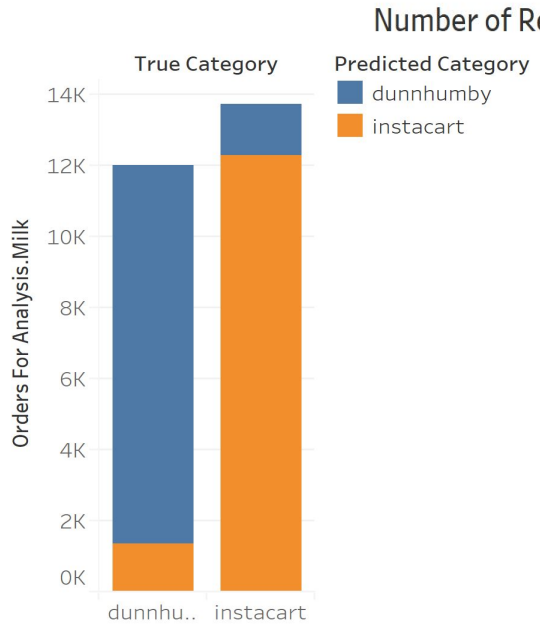
Machine Learning

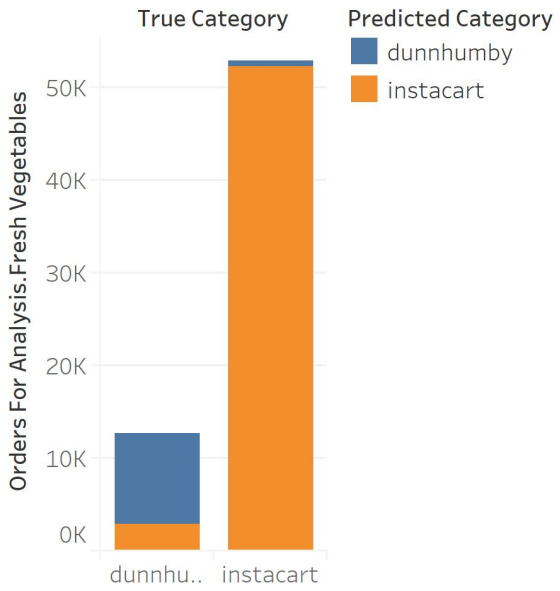
- Scikit Learn
  - Pyhive to access data from hive
- 108 features
  - Aisle categories
  - Day of week
  - Time of day
- Unbalanced classes
  - 220K Dunnhumby examples
  - 3 M Instacart examples
- Naive Bayes classifier
  - 86% accuracy

Grocery Shopper Classifier

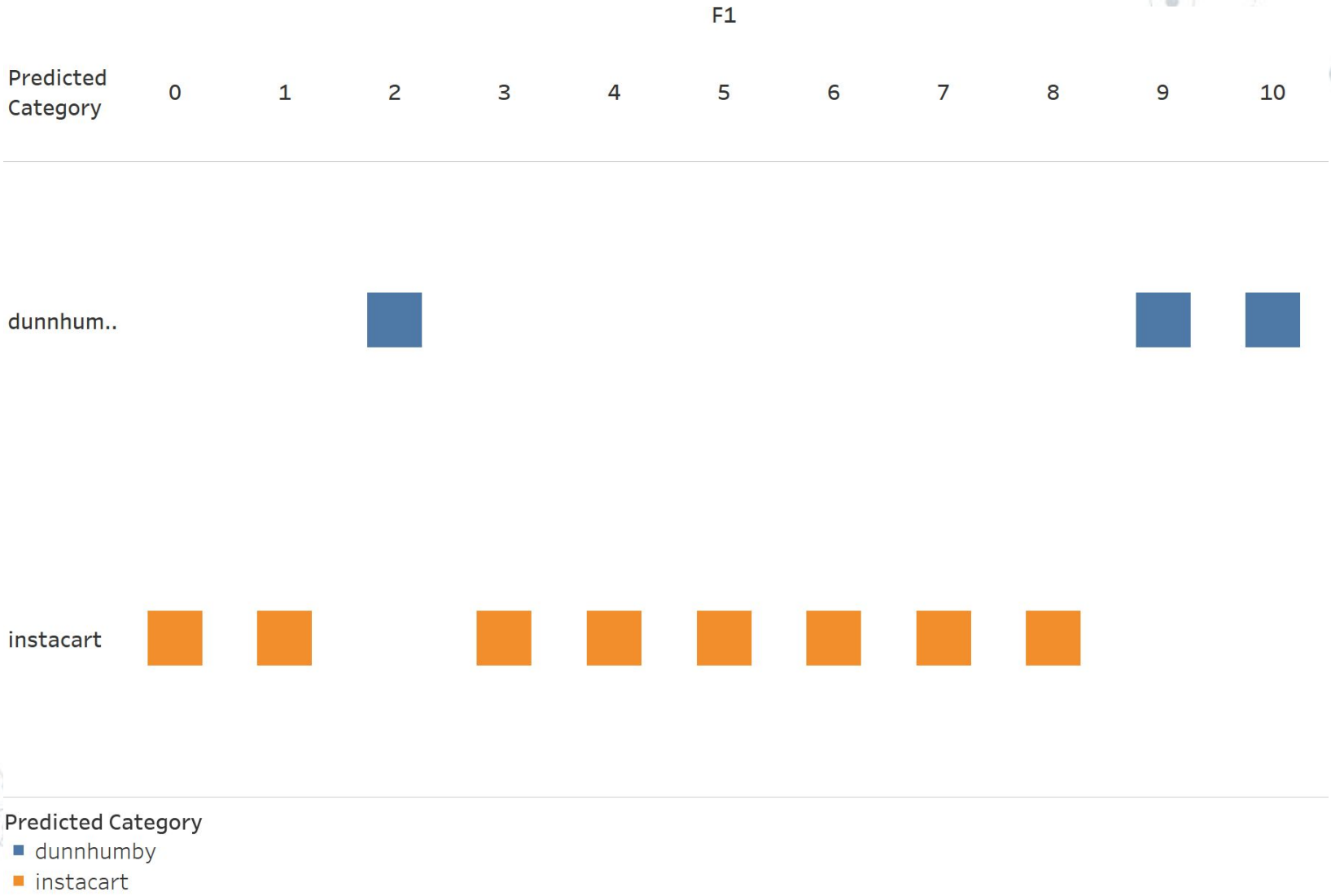








Grocery Shopper Classifier



Grocery Shopper Classifier

Demo

- SQL query

- Machine learning

Scaling Solutions

Current solutions

- Storage layer
  - Hive
- Processing layer
  - Hive SQL
- Visualization
  - Tableau
- Machine learning
  - Python with Scikit learn

Full scale out

- Processing Layer
  - Spark
- Machine learning
  - Spark MLlib

Future Plans How to Evolve Project

- Improve accuracy of shopper classifier
- Add a streaming layer to facilitate real-time shopper identification
- Tailor the shopping experience based on shopper category to increase sales
- Full scale out of the project