Carnegie Mellon University

# Project Progress Report

**MARCH 29, 2021**

Frank Kovacs, Ning Gao, Pragya Jain, Wonil Lee

# Agenda

- Introduction
  - Team Profile
  - Client Profile
  - Project Scope
- Dataset
- EDA
  - Issues Logged
    - Visualizations
    - Explanations
- Next Steps
- Q&A

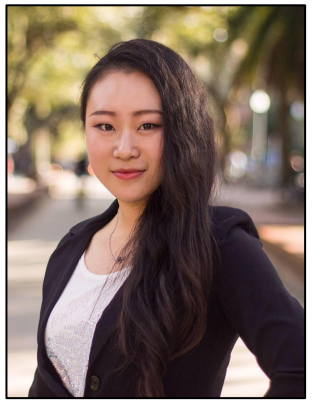
# Introduction

# Team

## Frank Kovacs

- CMU Statistics & Machine Learning ‘19
- Software & Data Research
- Research with Delphi COVIDcast and ISLE

## Ning Gao



- Georgia Tech Industrial & Systems Engineering ‘20
- Research with NSF LeapHi Program
- Past work experience in the telecom industry

## Pragya Jain

- Past work experience in the insurance industry
- Associate Actuary
- B.E. from NSIT, New Delhi

## Wonil Lee

- Past work experience in Consulting (2+ years)
- CMU Tepper & Statistics ‘18
- R, SQL, and Python

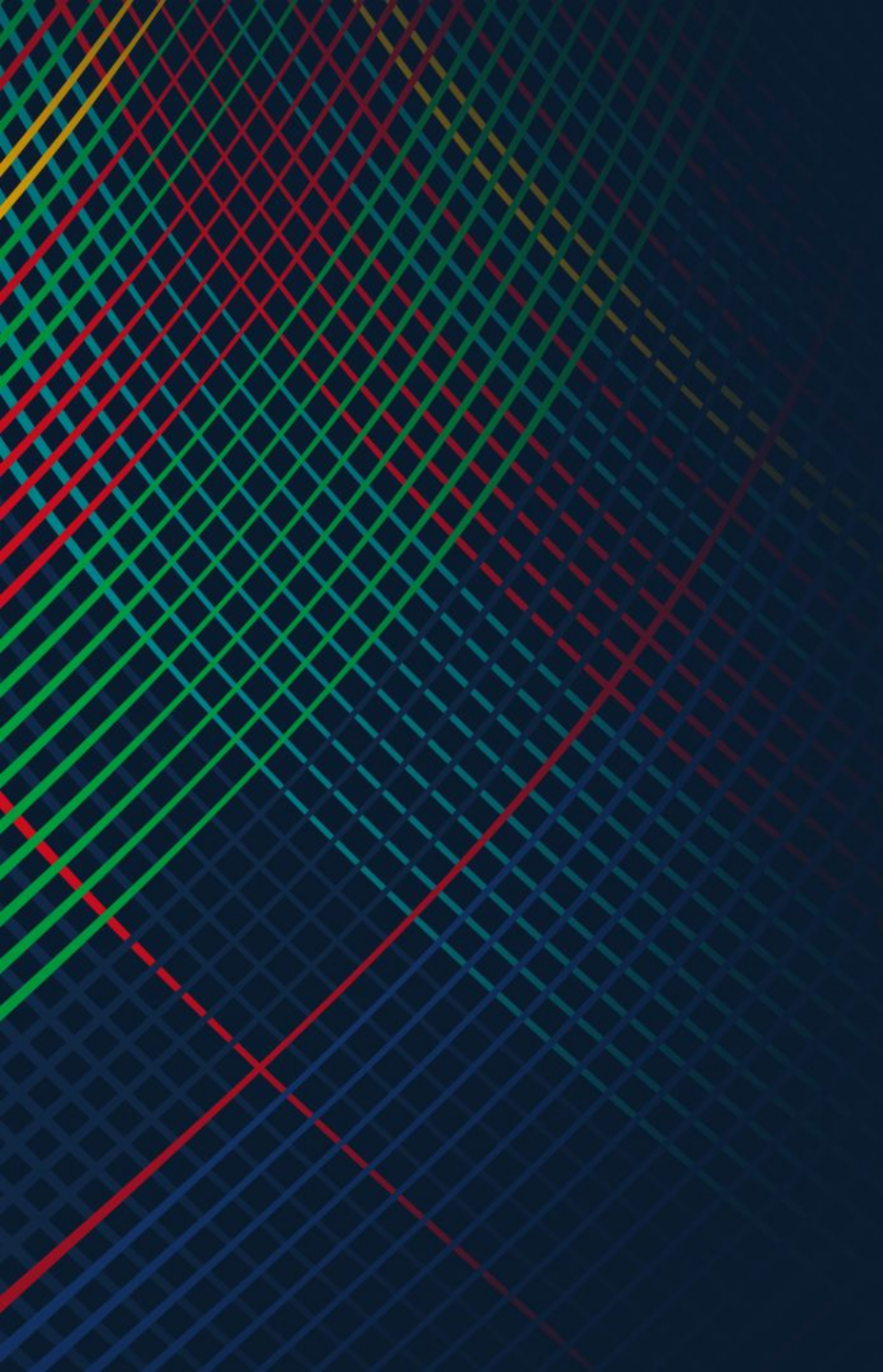
# NPD Group Overview

- NPD Group is a **Market research company**
- “Raw data assets into insights”
- Specialize in general merchandise and food service
- Market leader
  - **8B+** B2B transactions / yr

# Objective & Scope

- "…explore using unsupervised learning methods to help identify common data collection errors to help guide further analyst review."
- **Goals**
  - **Detect anomalies in time series datasets**
  - Identify common data collection errors
  - Facilitate further data analyst review
  - Automate data error flagging processes

 *Source: Capstone Handout*



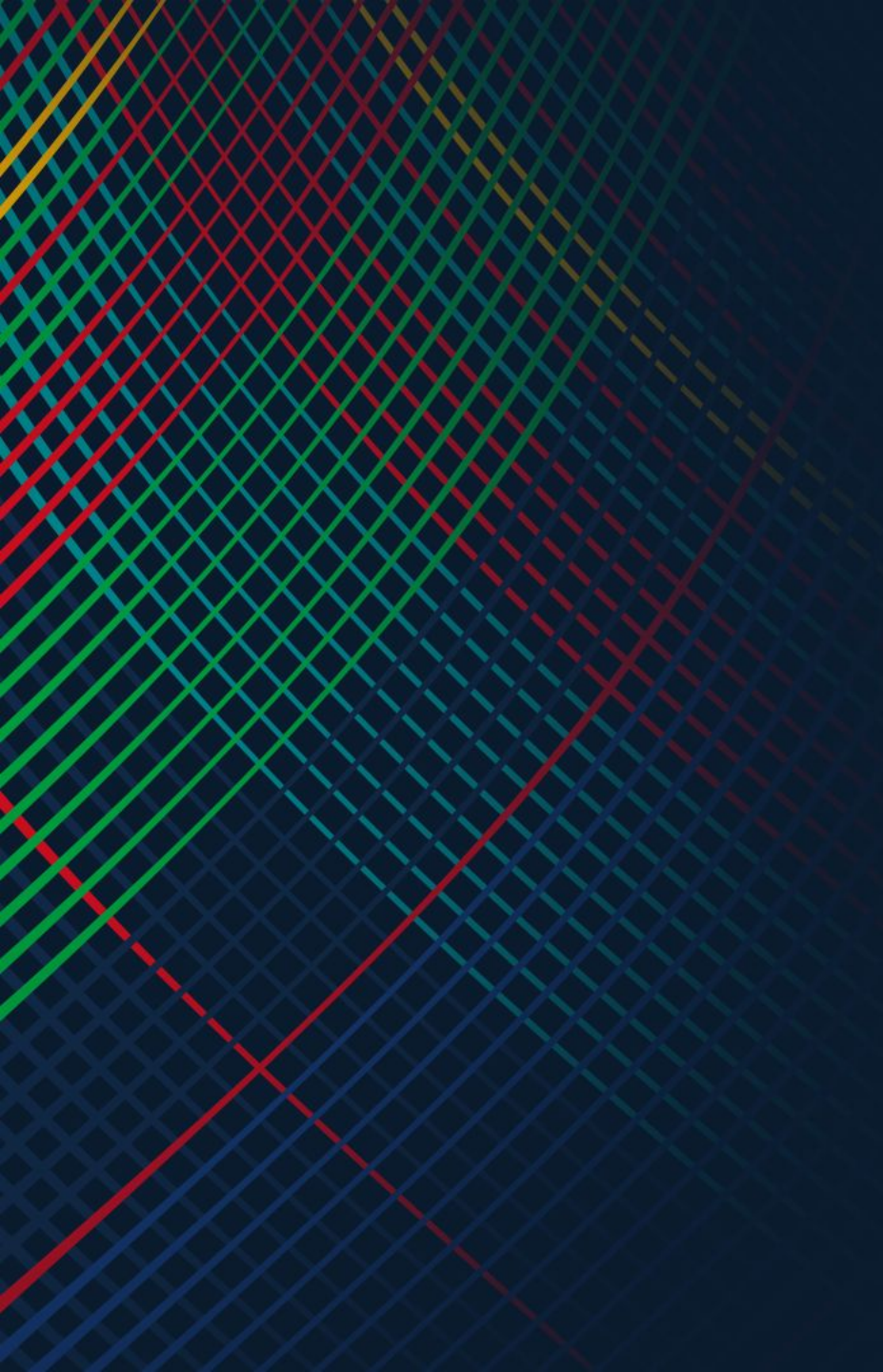
# Dataset

# NPD Project Dataset Overview - Key Variables

- **Merchant ID and Name**
- **Acquire Type ID**
- **Receipt_count**
- **Sum_total_paid**
- **Item_total**
- **Sum_items_distinct**
- **Sum_item_spend**
- **Panelists**

# NPD Project Dataset Overview - Main Datasets

- **Source Data**
  - 516 rows, 8 columns
  - Weekly values of the receipt_count, sum_total_paid, sum_items_distinct, sum_item_spend, panelists by 4 different data source types (iPhone, Android, Sift, and Receipt pal on device)
- **Retailer Data**
  - 983,953 rows, 11 columns
  - Weekly values of the receipt_count, sum_total_paid, sum_items_distinct, sum_item_spend, panelists by individual merchants and by different data sources
- **Issue Data**
  - 31 rows, 5 columns
  - Dataset of when (the Acquired date) and where (merchant name & source type) the data collecting error occurred



# EDA

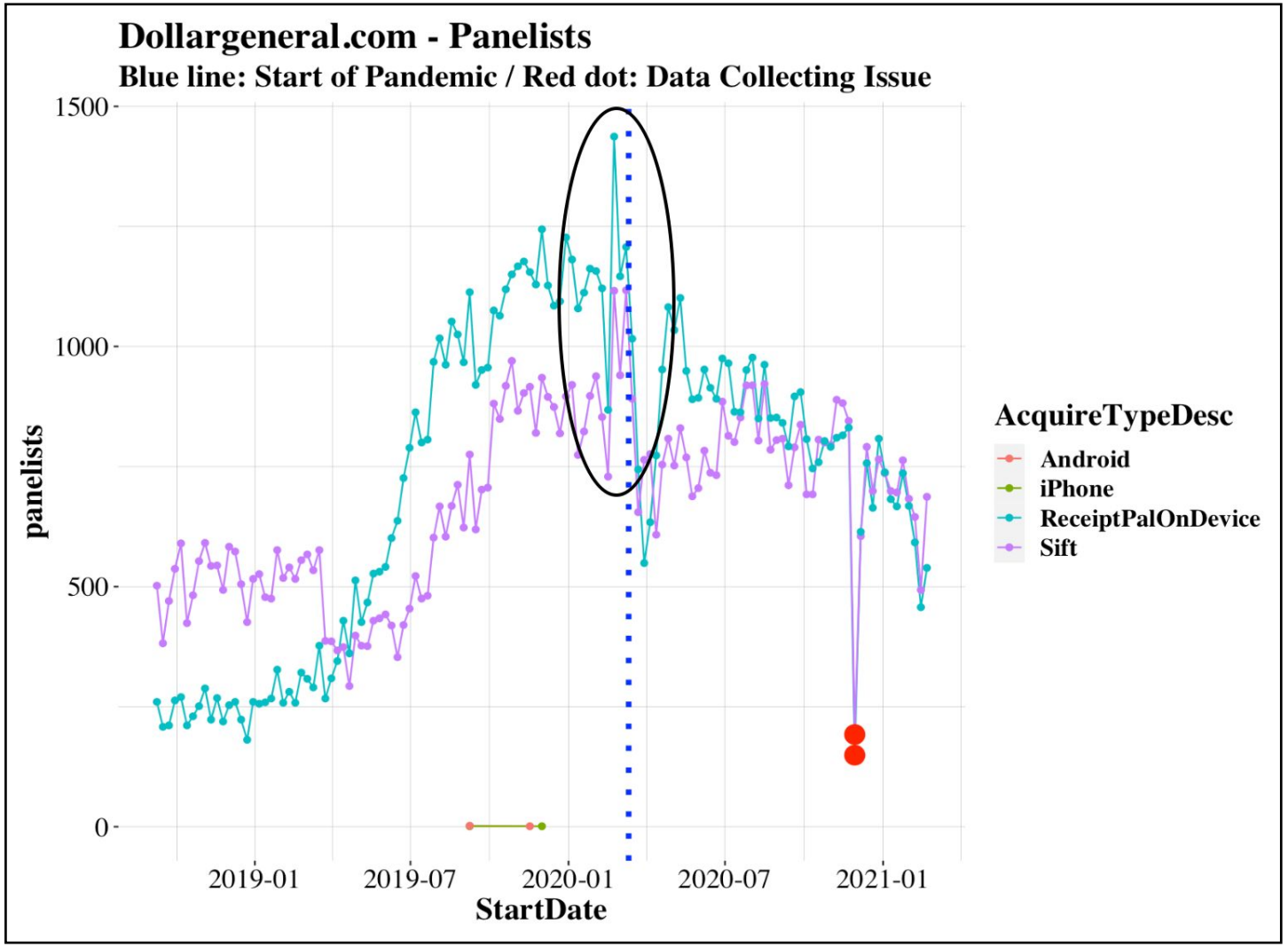
# EDA

- **Existing Flags**
  - Issues logged by client in the past 2 years were shared
- **Data Preparation**
  - Data sanity checks
  - Merged ‘Retailer Data’ with ‘Issue Data’
- **EDA Plots**
  - Generated time series visualizations for individual merchants and marked issues logged by client with a ‘Red Dot’
  - Start of Pandemic marked with a vertical line - March 11th, 2020
  - Highlighted potential unmarked anomalies

# Issues logged | Visualization Explanation

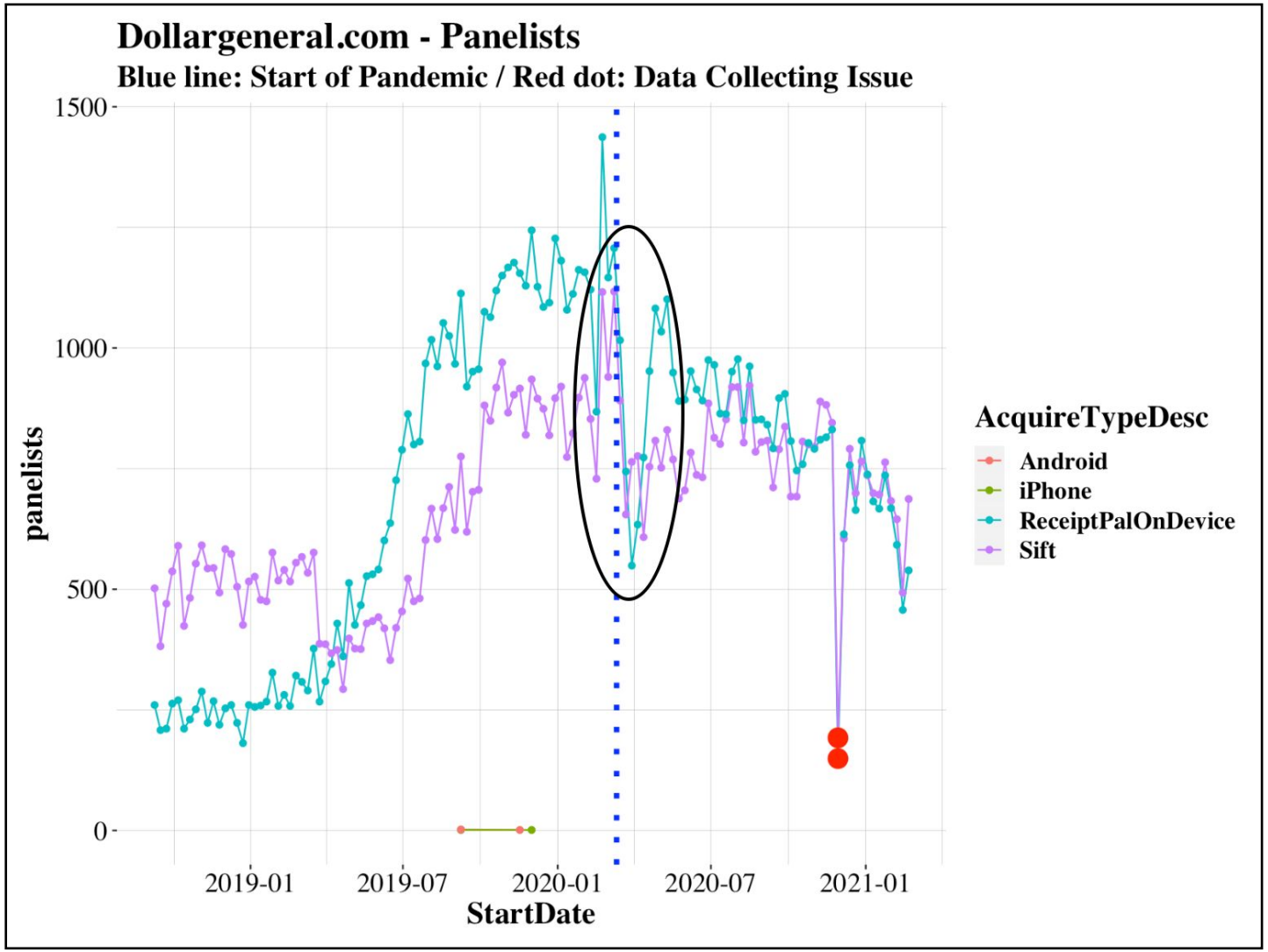
# Preference for marking dips as issues

- A peak of significant amplitude has not been marked as an issue
- A dip of similar amplitude has been marked as an issue



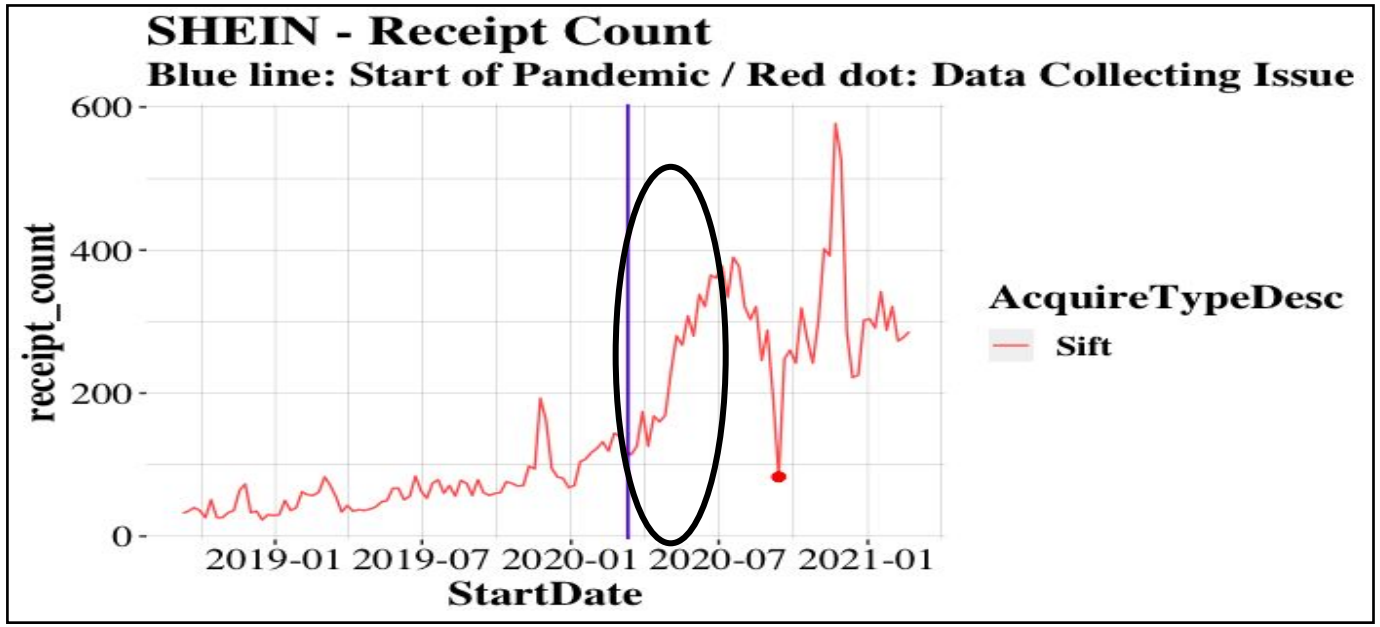
# Anomalies right after Pandemic Start not marked

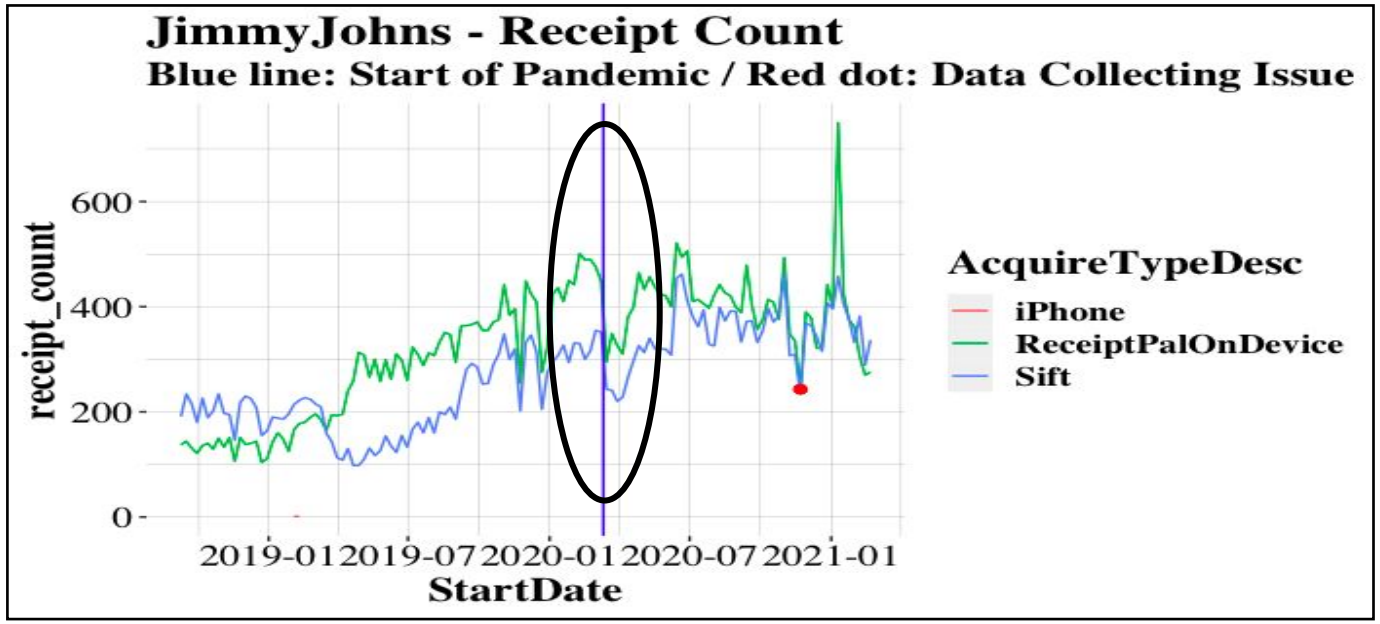
- Detected drop near Christmas 2020
- Did not detect drop near start of Covid-19



# Anomalies right after Pandemic Start not marked

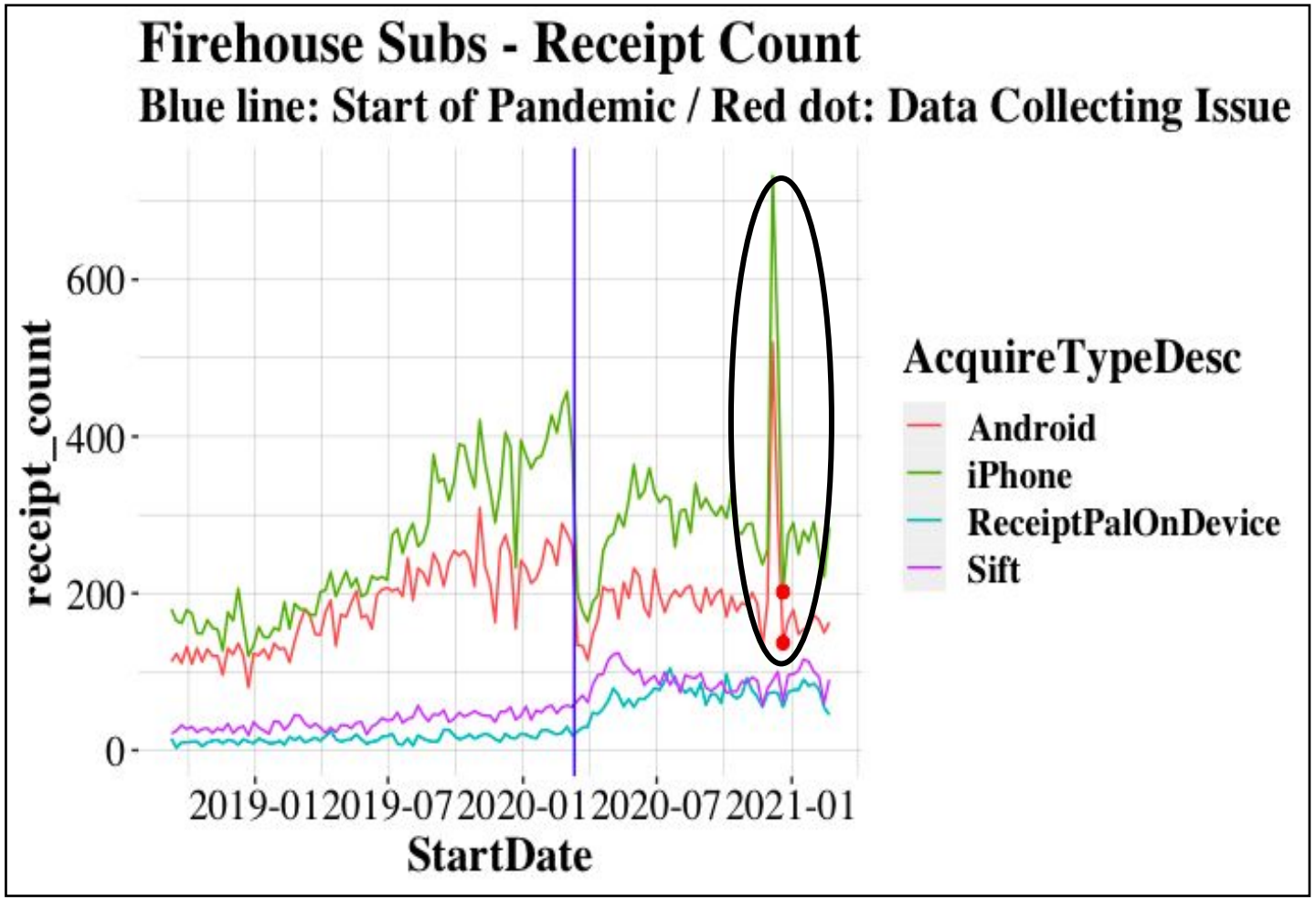
- The anomalies after the Pandemic outbreak is not captured
  - Sudden Decrease
  - Sudden Increase
- We would like to know whether the current algorithm considers impact of COVID19



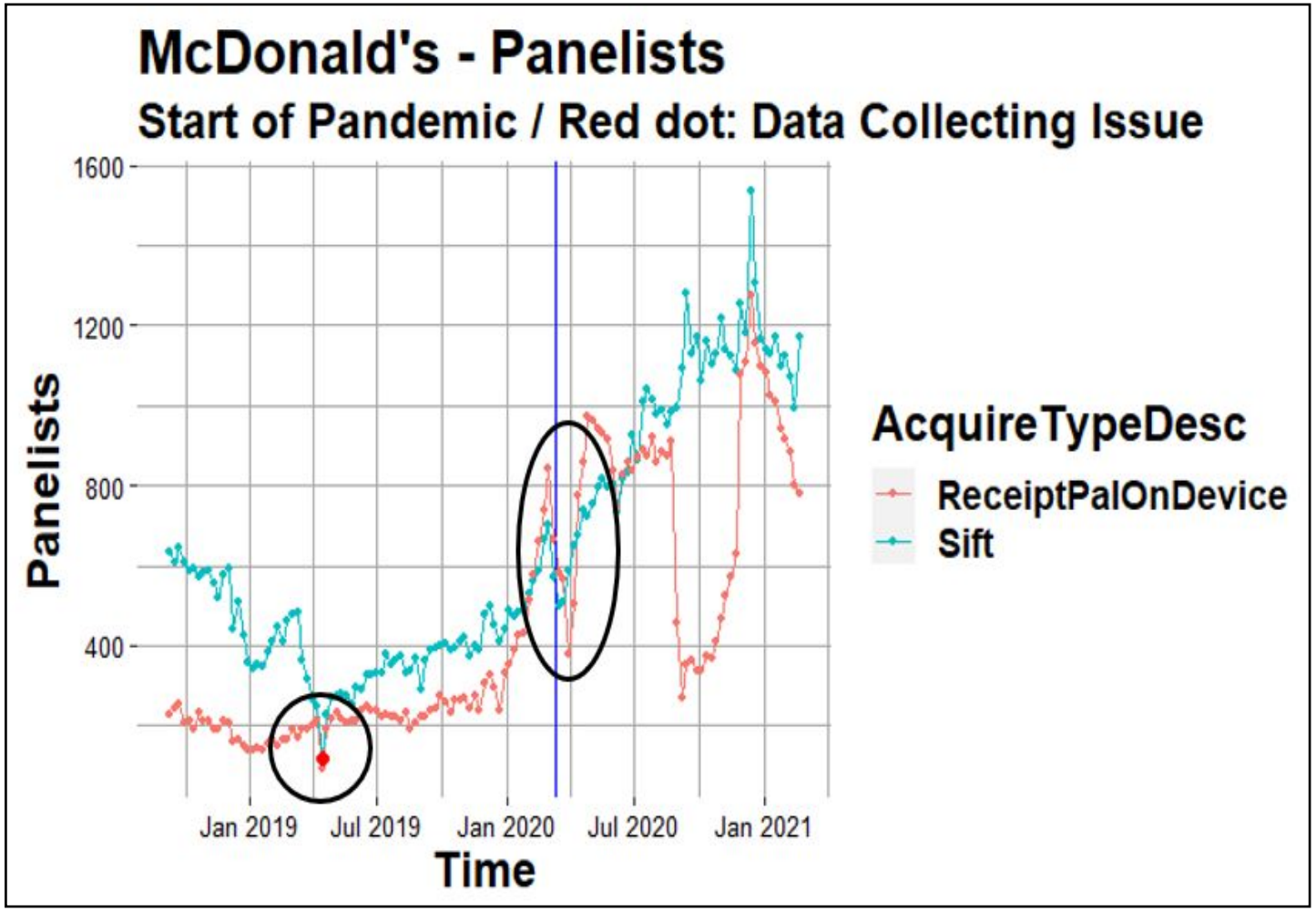
# Delayed detection of Sudden Shifts

- Anomalies are detected in delayed manners
  - The error was detected 3 weeks after the first abnormal value



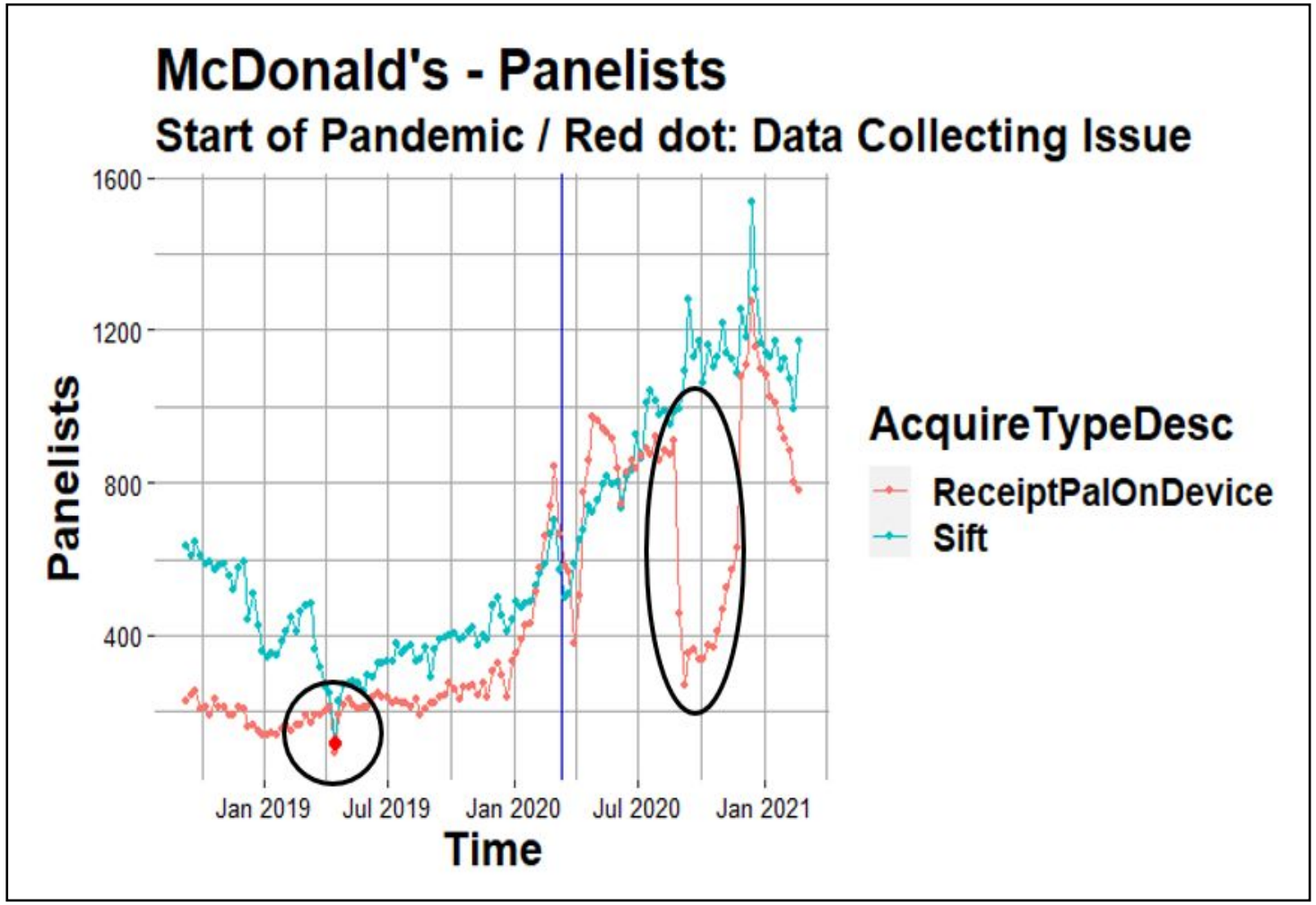
# Detection of rapid dips over 2-3 weeks but not over 1 months

- Small, sharp drop of panelists using Sift around April 2019
- Bigger, consistent 1 month drop around pandemic time
- Do we want to detect long-term anormales?



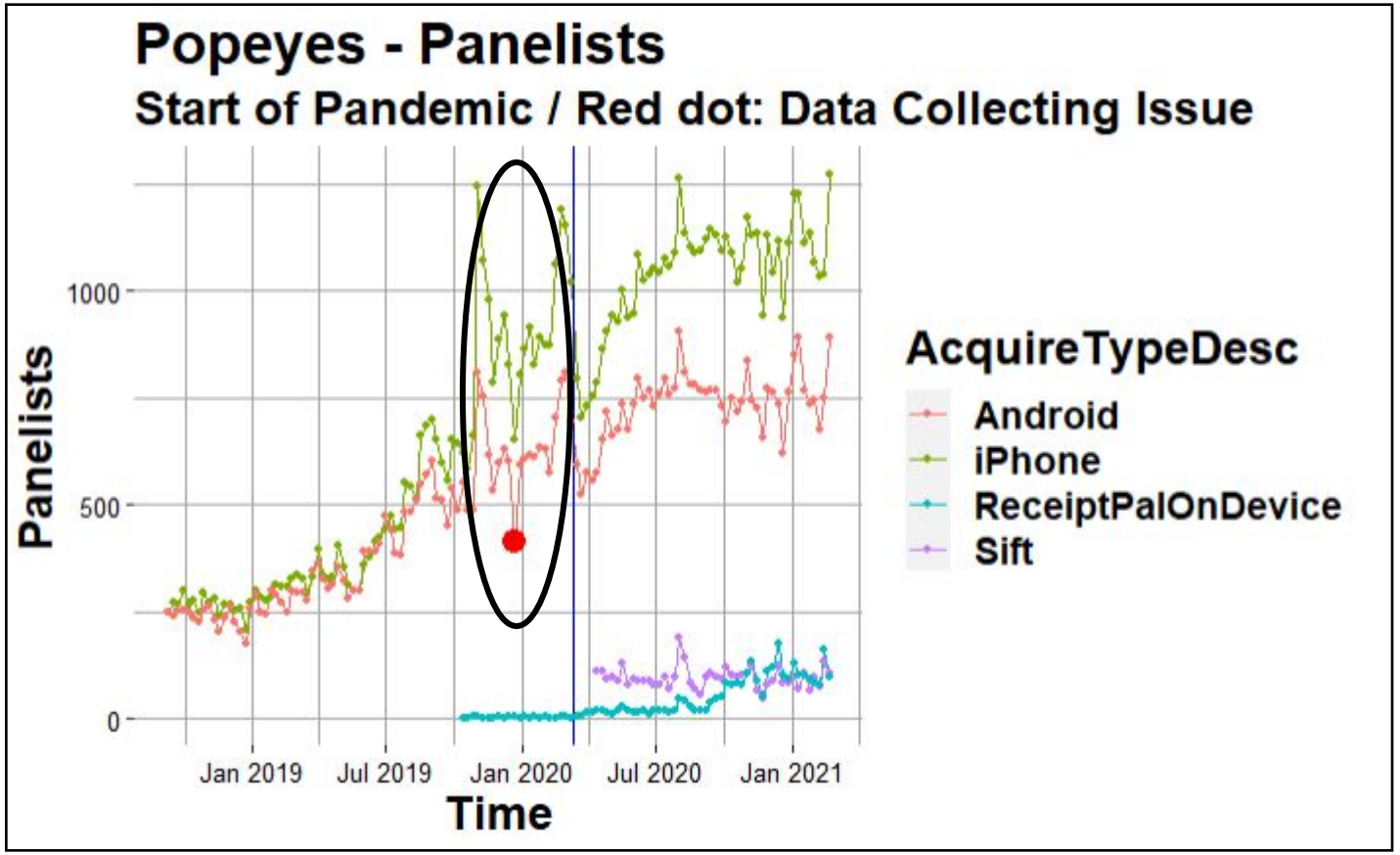
# Smaller Amplitude of dips detected but not bigger ones

- Small, sharp drop of panelists using Sift around April 2019
- Huge, sharp drop from September to November 2020
- Drops were similar, but one is detected, the other is not



# Dips for all Acquire-Types not marked

- Dip in Android was marked as an issue
- Dip in iPhone for a similar time period and amplitude, but was not marked



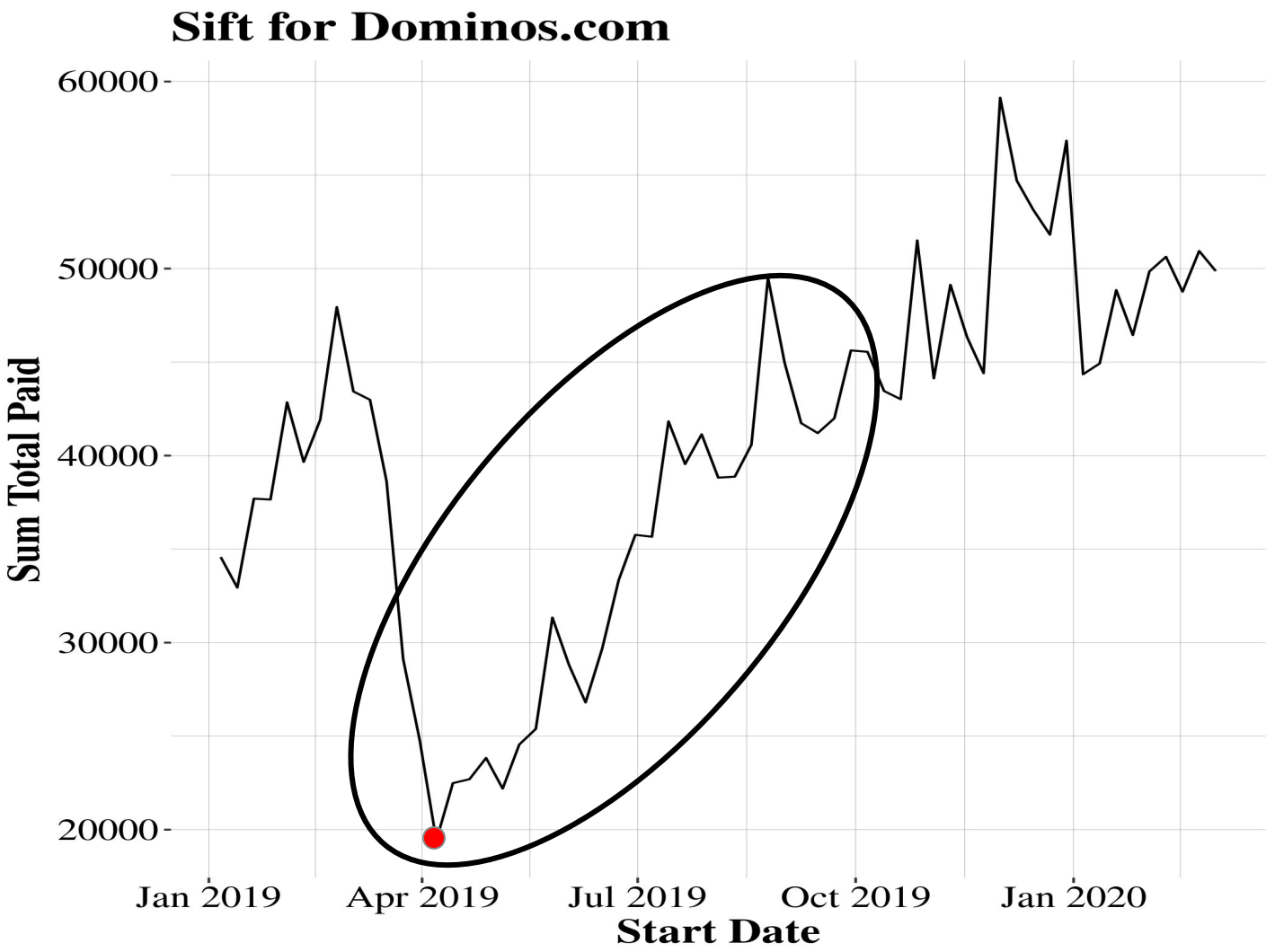
# Missing data - Jersey Mike's

- Issues flagged for the dates that did not have data
- No definition given, assume missing data

| Date | Retailer Data Present? | Issue Found? |
|---|---|---|
| 7/19/20 | Yes | No |
| 7/26/20 | No | Yes |
| 8/2/20 | No | Yes |
| 8/9/20 | No | Yes |
| 8/16/20 | Yes | No |
| 8/23/20 | Yes | No |
| 8/30/20 | Yes | No |
| 9/6/20 | No | Yes |
| 9/13/20 | No | Yes |
| 9/20/20 | No | Yes |
| 9/27/20 | No | Yes |
| 10/4/20 | Yes | Yes |
| \vdots | \vdots | \vdots |

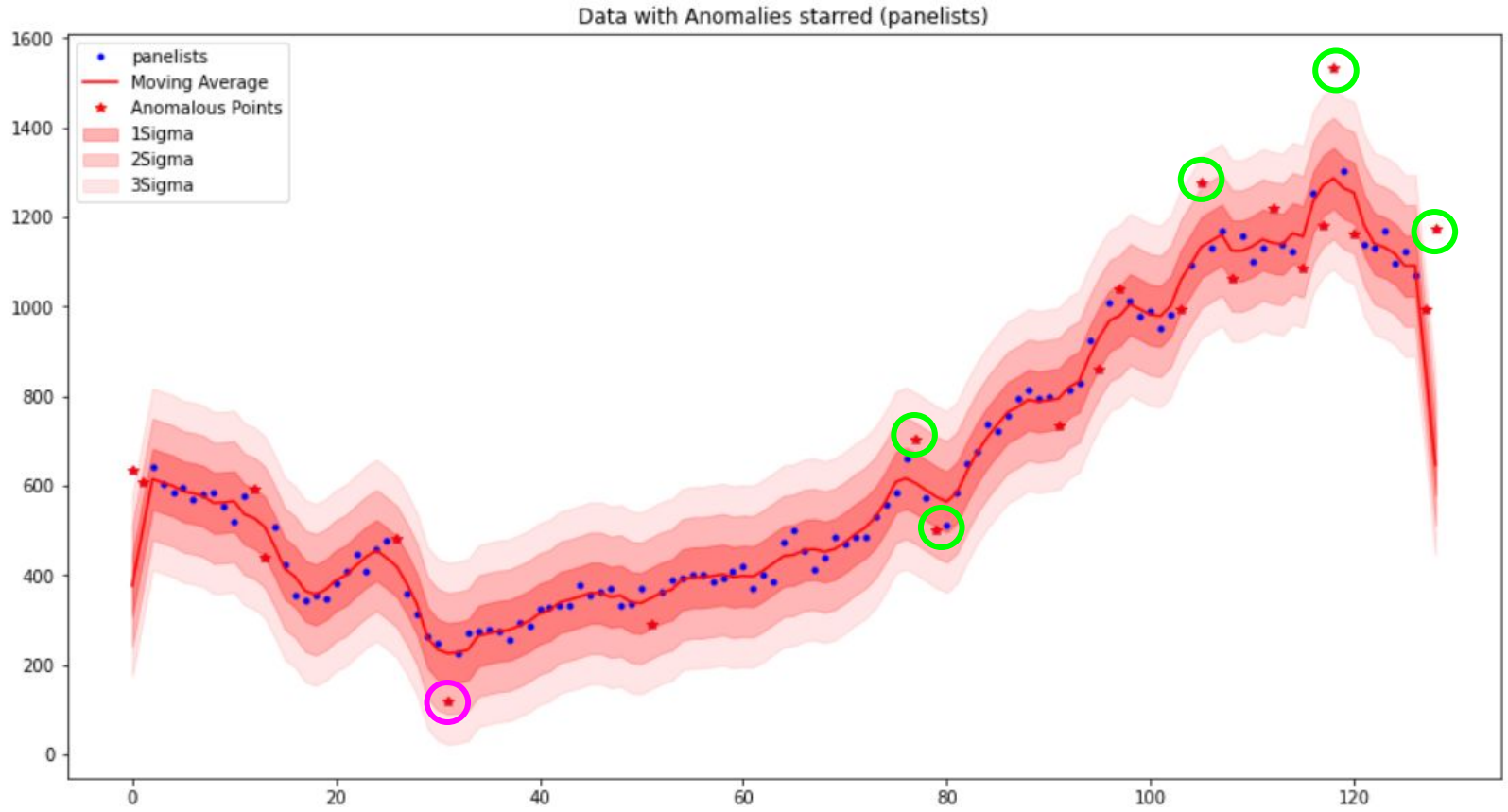
# Should Trends be Flagged?

- Many of the graphs had clear trends (slow drift) that were not detected.
- We see a shift starting from April 2019 that increases until approximately November 2019
- Should we detect trends? If so, what kind of trends?



# Current Approach

# Moving Average Prototype



# Moving Average Prototype

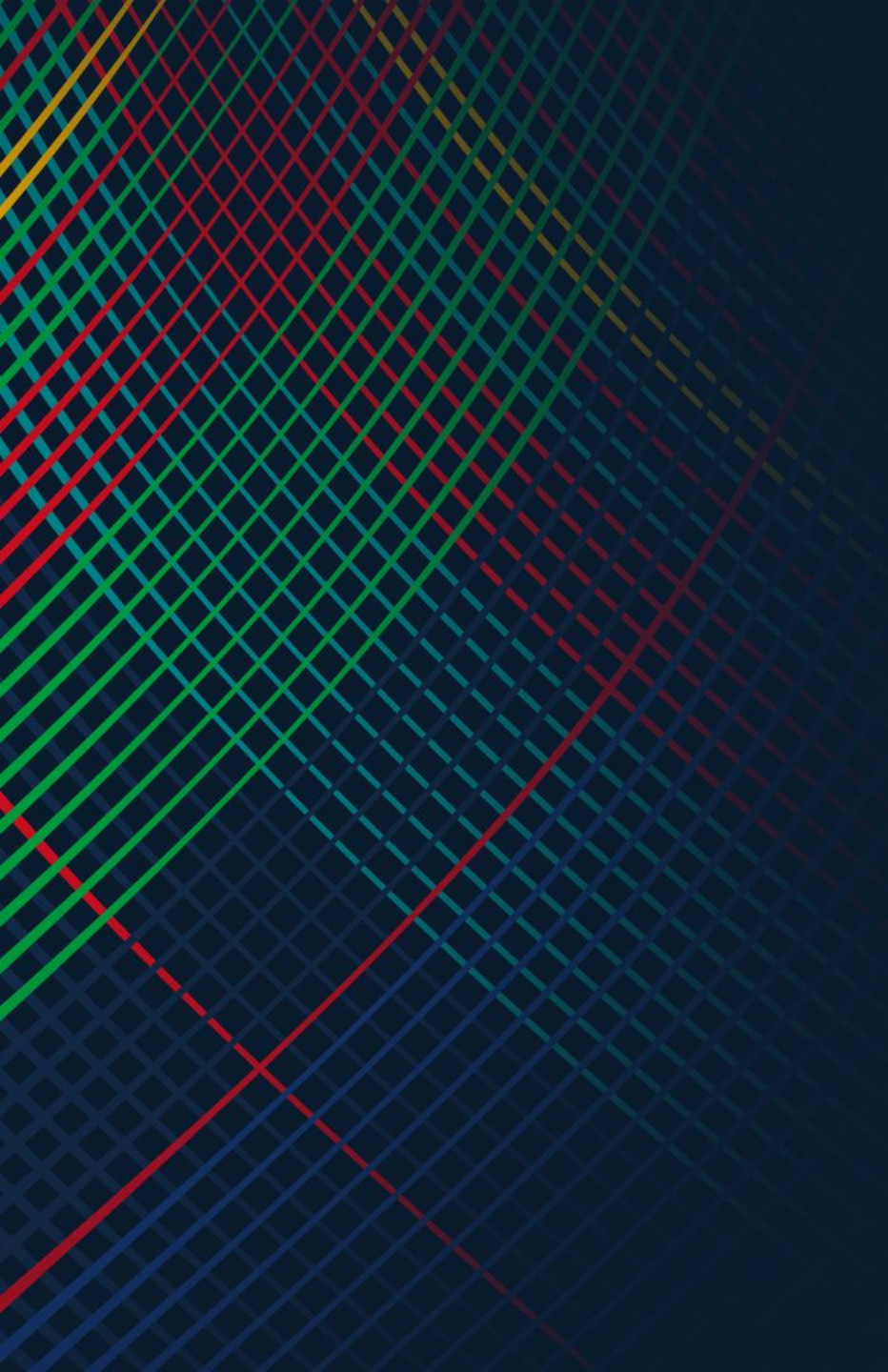
- Given window, standard moving average
- Differences attributed to variation in panelists
- Problem = do not have individual-data, only have weekly sums
- Solution = modified MLE estimation of parameters
  - 1. Exploit distribution of panelists
  - 2. Scaling of sum variables

* Technical documentation available on request *

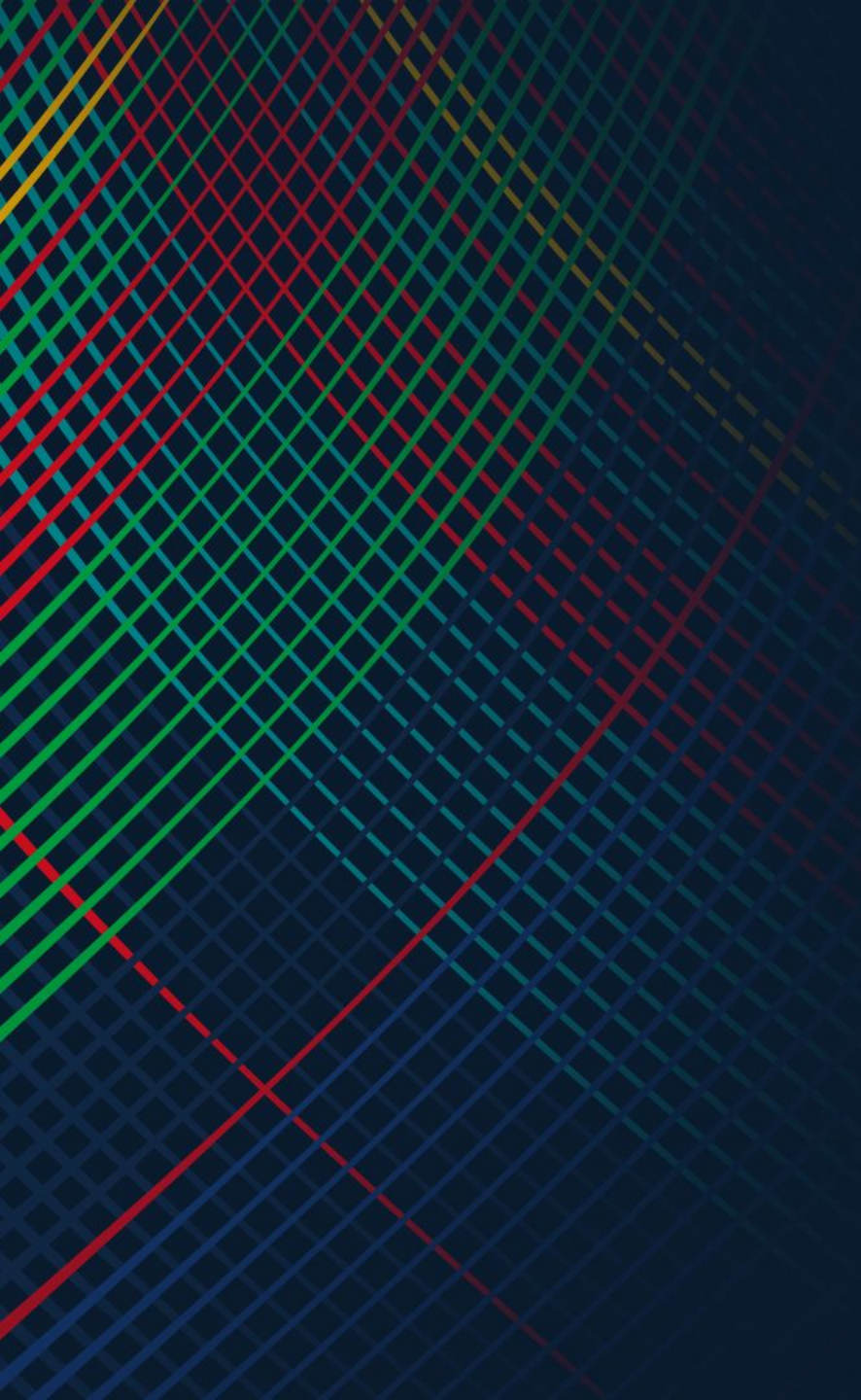
# Next Steps

# Next Steps

- Currently developing the prototype algorithm of anomaly detection
  - MA model with sliding window
  - MLE method
- Client meeting tomorrow (March 30, 2021)
  - Receive feedbacks on EDA and group's follow-up questions
  - Share the approaches used for development of the prototype algorithm with the client
- Further research on anomaly detection methods in a time series
  - Facebook
  - Twitter



# Q&A



# THANK YOU!