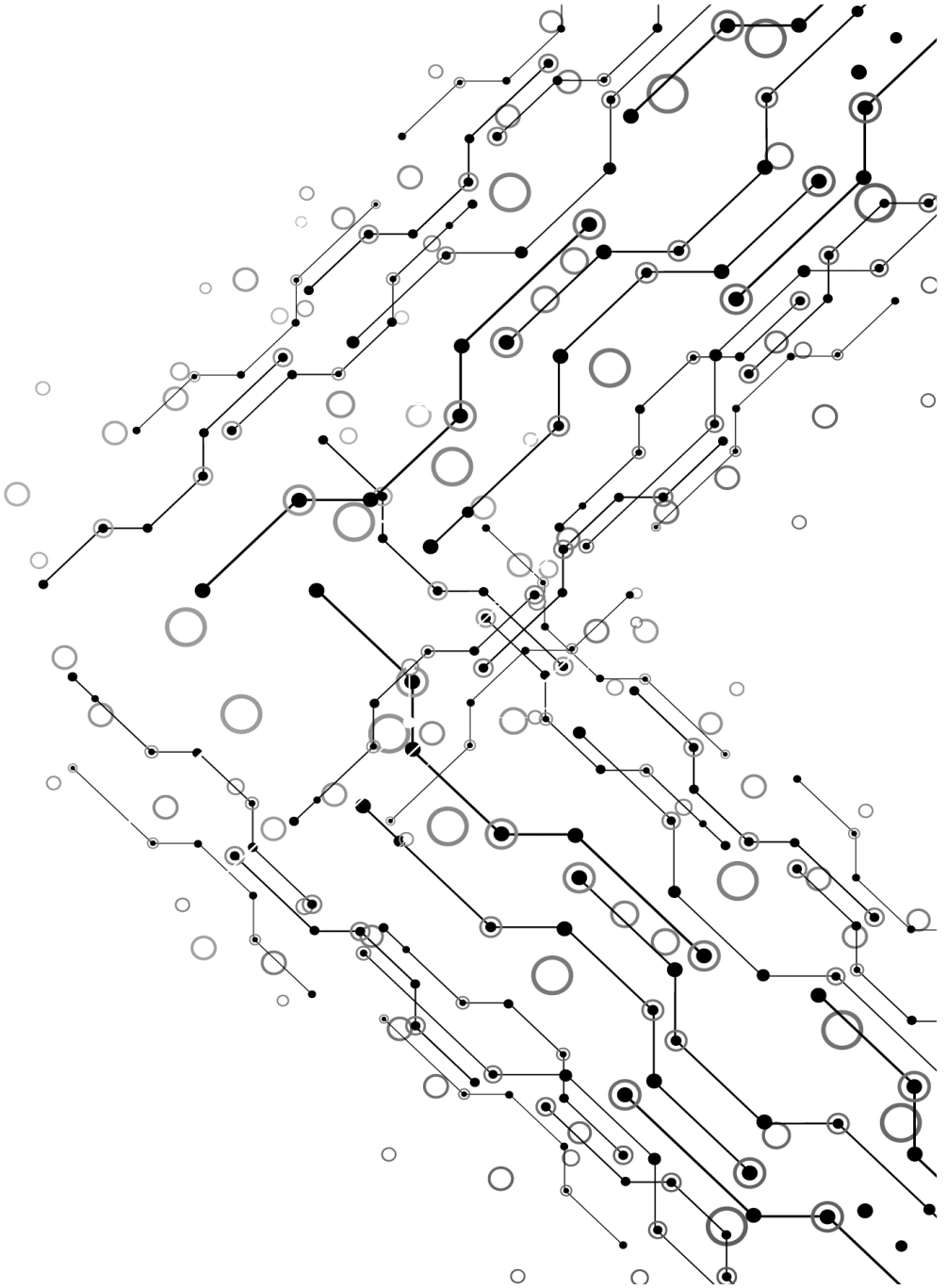
Food Consumption Trend Detection Report

—Based on Facebook text data

Aidi Zhang, Marketing Analytics



Summary

This Facebook text analysis is mainly based on the **term frequency of document-term matrix.**

By exploring the term frequency trends of the two existing topics (cauliflower rice and vegetable noodle), I developed a set of methodology to detect increasing co-occur words and the emerging food consumption trend in the early stage. The analysis process and result will be displayed in the rest of this report.

Analysis Process:

- Explored term frequency trends across time for two existing topics.
- Explored term frequency proportion trends across time for two existing topics.
- Plotted moving average trend with actual trend of two topics.
- Wrapped up methodology.

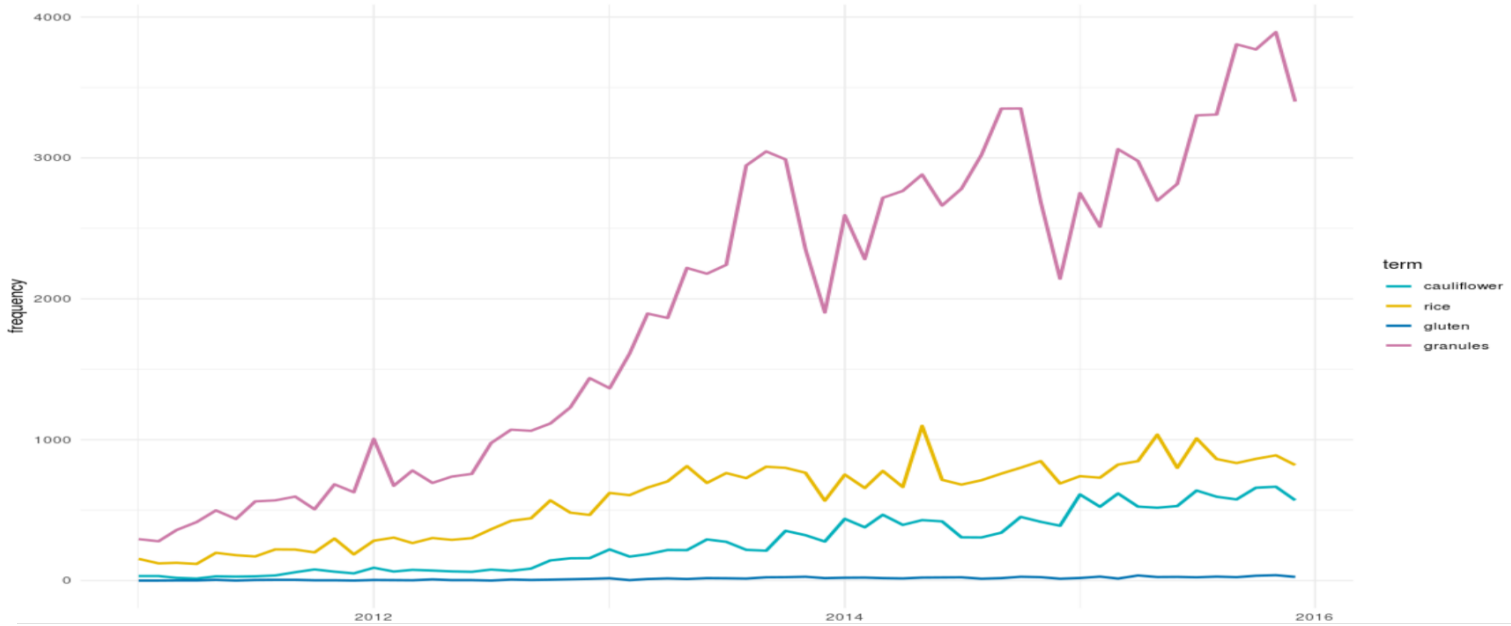
Methodology:

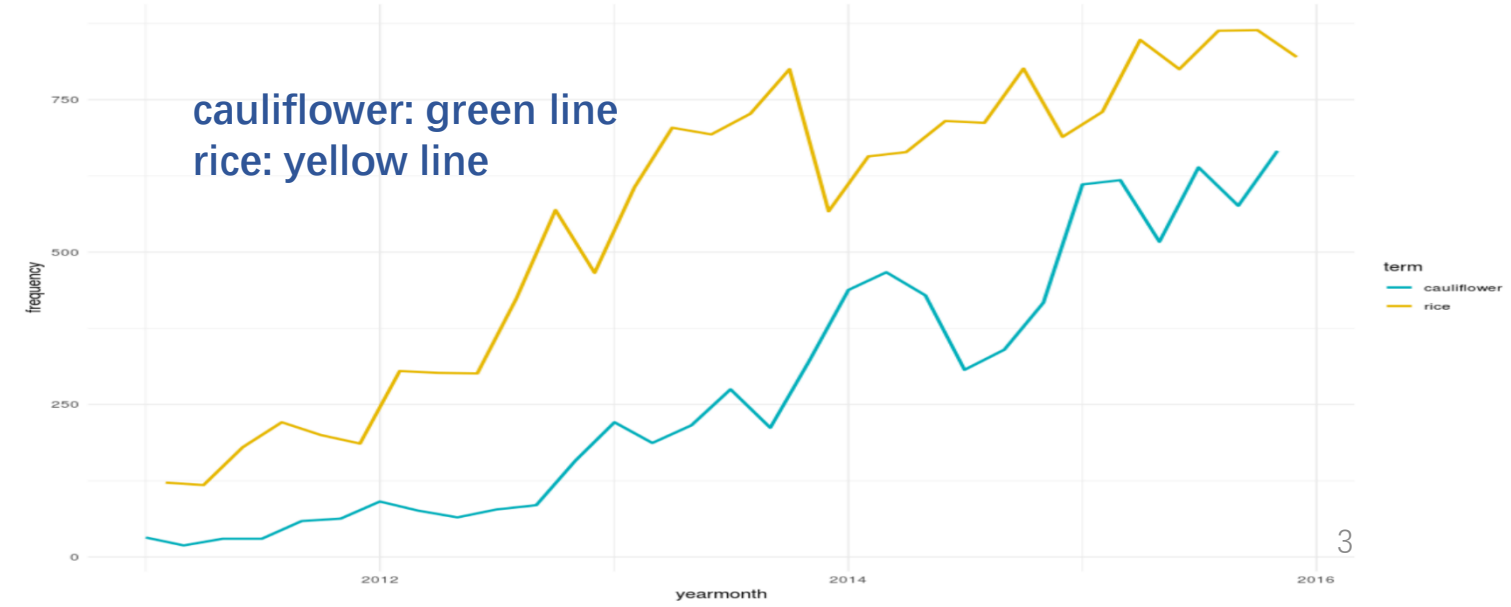
- Term Frequency Table – DTM
- Moving Average Model
- Visualization with ggplot

Analysis Demonstration for Cauliflower Rice Topic

Part 1. Time Series *Term frequency* trends for cauliflower rice.

- I picked the words which are relevant to "cauliflower rice": **"cauliflower", "rice", "gluten" and "granules"** from the introduction. After some data cleaning and manipulation, I got the frequency line chart.
- The words "cauliflower" and "rice" share similar patterns of overall increase in their frequency across time.
- This aligns with our result that "cauliflower rice" consumption has emerged over time and has become a real trend in **2016**.

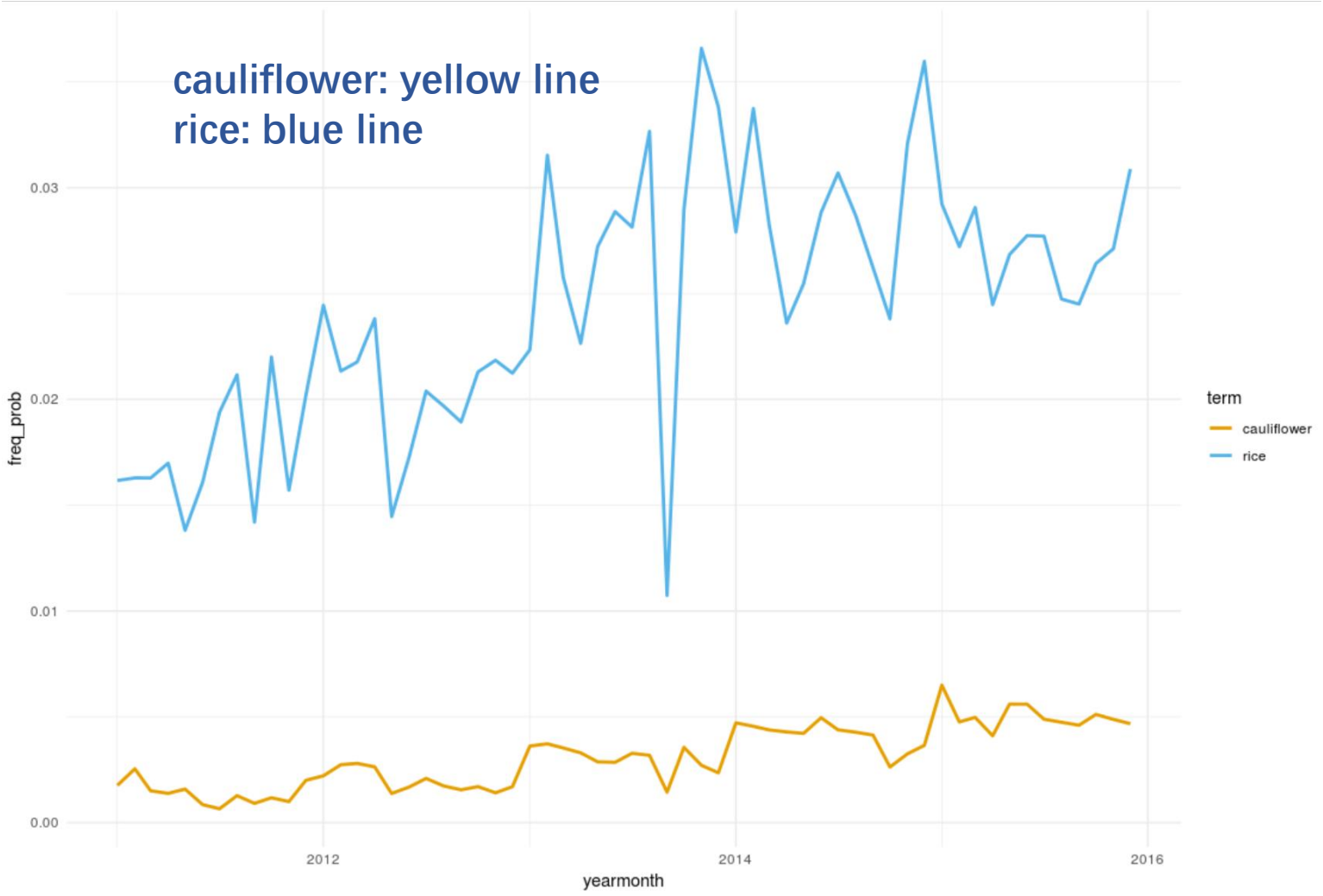




Analysis Demonstration for Cauliflower Rice Topic

Part 2. Time Series *Term frequency proportion* trends for cauliflower rice.

- **"Rice"** accounts for **larger proportion** in the whole corpus in five years, and the pattern **fluctuates** more.
- I speculated that there are more **false positives** in word 'rice", which means many occurrences of "rice" didn't associated with cauliflower.
- Therefore, "**cauliflower"** is the more accurate representation of the "cauliflower rice" topic, and I further explored this word frequency in the next page to facilitate emerging trend detection.



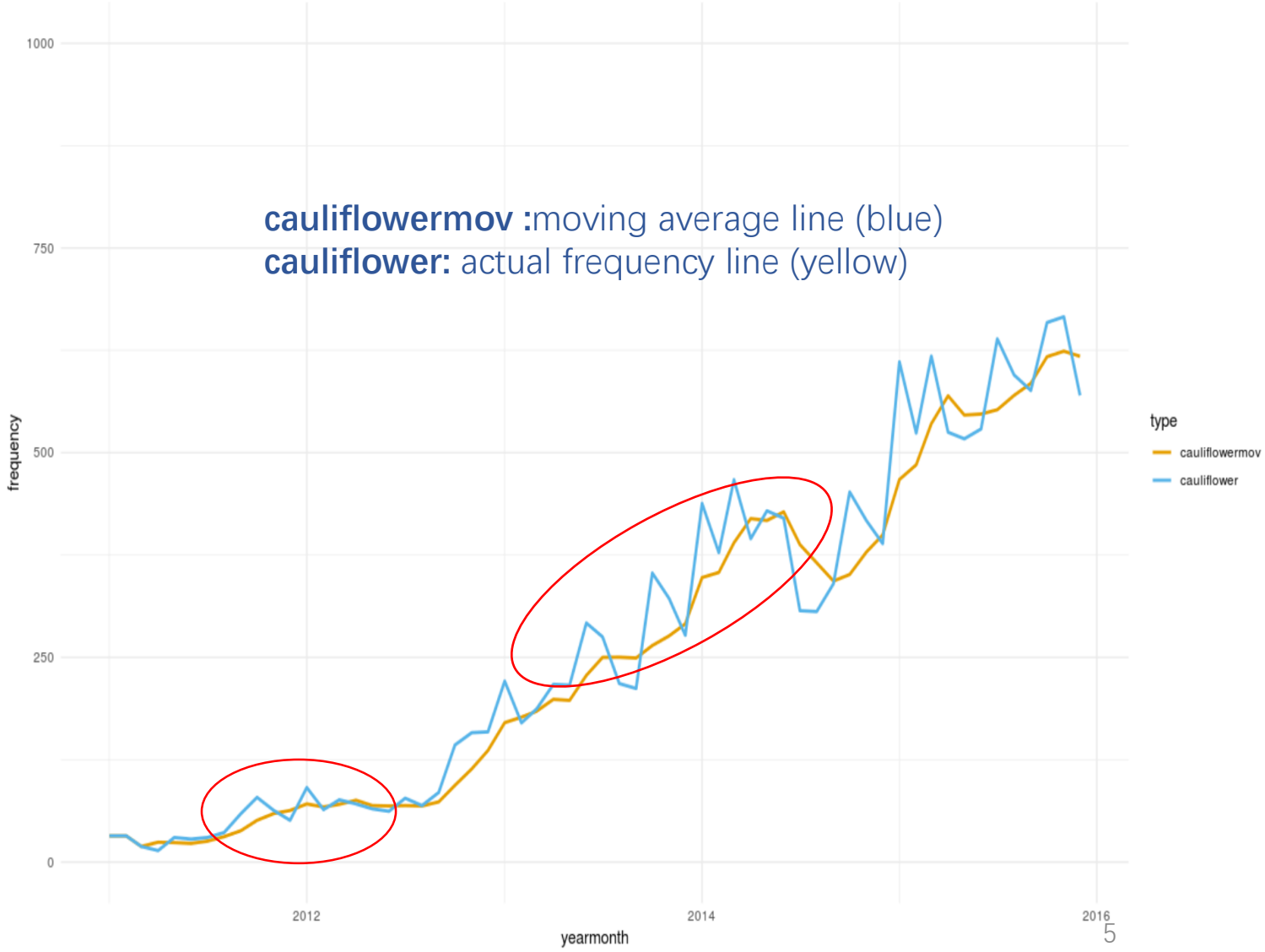
Analysis Demonstration for Cauliflower Rice Topic

Part 3. *Moving Average* frequency trend vs. Actual trend of the word "cauliflower".

- From the **late-mid of 2011**, the moving average frequency line of "cauliflower" has started to show a trend to be above the actual frequency line.
- From the **end of 2013**, the trend has started to become more and more obvious, and there was several shot-term sharp increases of the topic around **2014**, then it became really trendy in **2015**, and in **2016** the trend has been followed by the public.

Speculation:

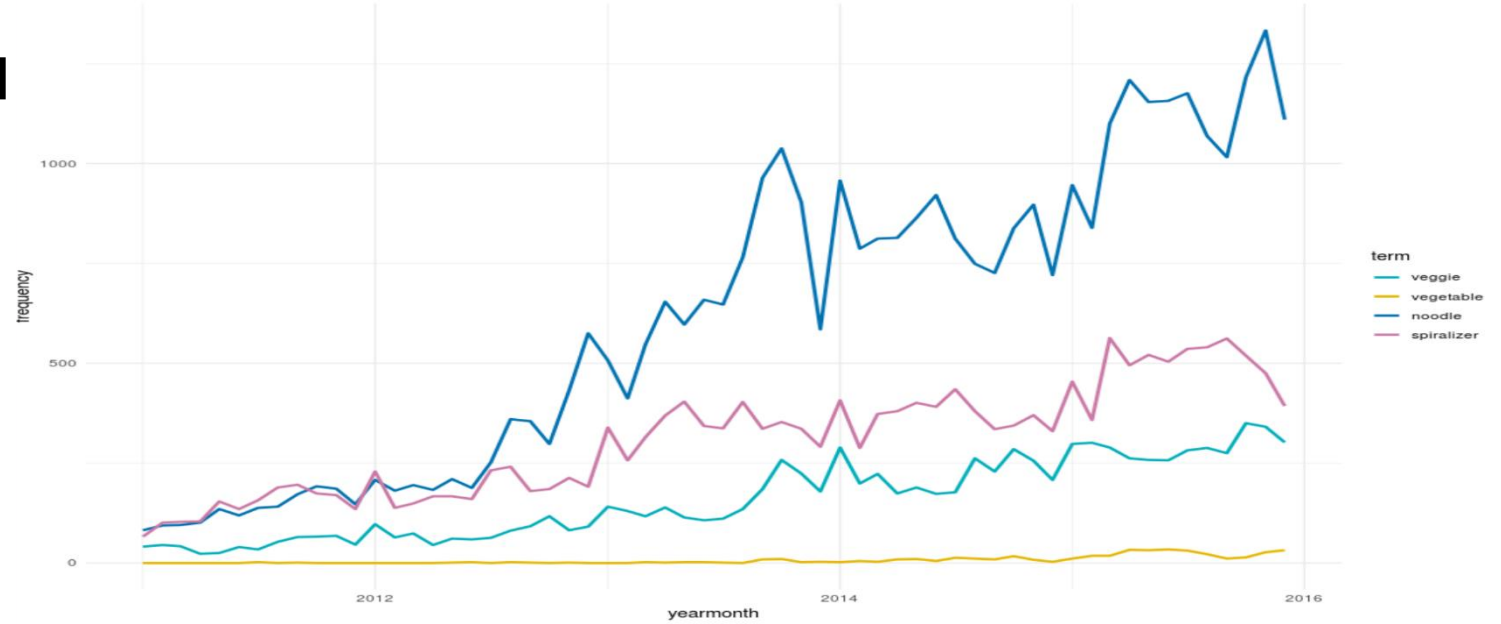
- Based on this chart, we can detect the trend as early as the **late-mid 2011**, and we can validate it after **2013**.

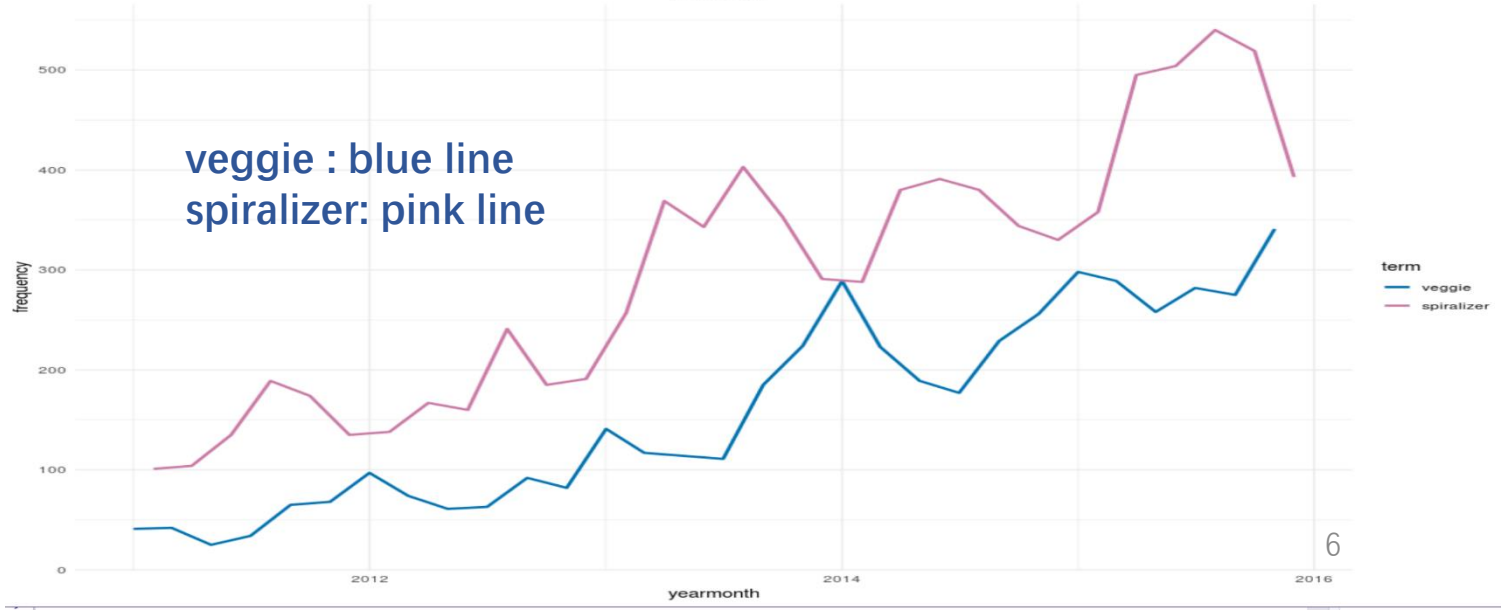


Analysis Demonstration for Vegetable Noodle Topic

Part 1. Time Series *Term frequency trends* for vegetable noodle.

- I picked the words which are relevant to "vegetable noodle": **"veggie", "vegetable", "noodle", "spiralizer"** from the introduction. After some data cleaning and manipulation, I got the frequency line chart.
- The words "veggie" and "spiralizer" share quite similar patterns of overall increase in their frequency across time. The line of "spiralizer" arrived at a peak in the **middle 2015,** and has dropped a little bit since the **end of 2015**.
- This aligns with our finding result that "vegetable noodle" consumption has emerged over time and has become a real trend in **2015**.

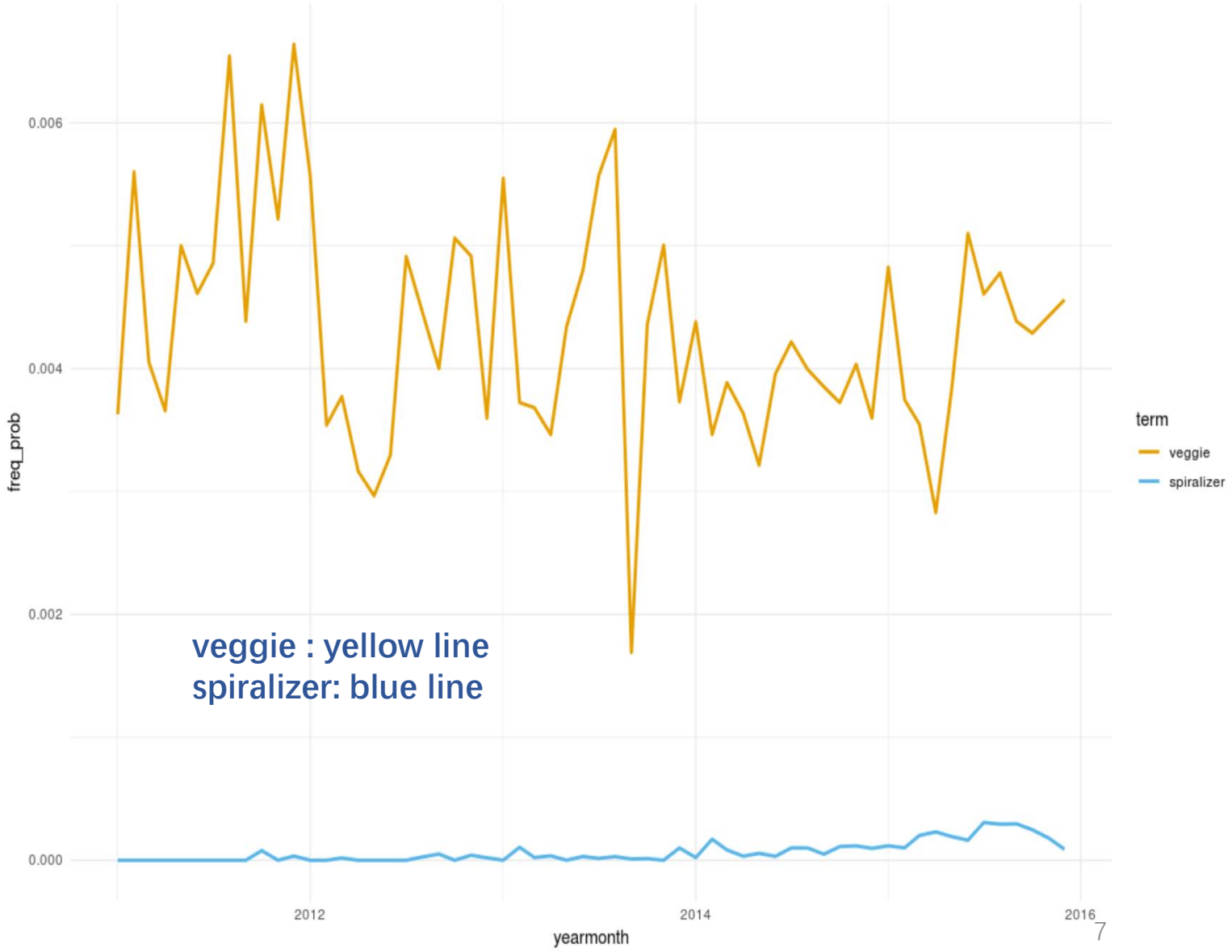




Analysis Demonstration for Vegetable Noodle Topic

Part 2. Time Series *Term frequency proportion* trends for vegetable noodle.

- **"Veggie"** accounts for **larger proportion** in the whole corpus in five years, and the pattern **fluctuates** more.
- I speculated that there are more **false positives** in word 'veggie", which means many occurrences of "veggie" didn't associated with spiralizer.
- Therefore, "**spiralizer"** is the more accurate representation of the "vegetable noodle" topic, and I further explored this word frequency in the next page to facilitate emerging trend detection.



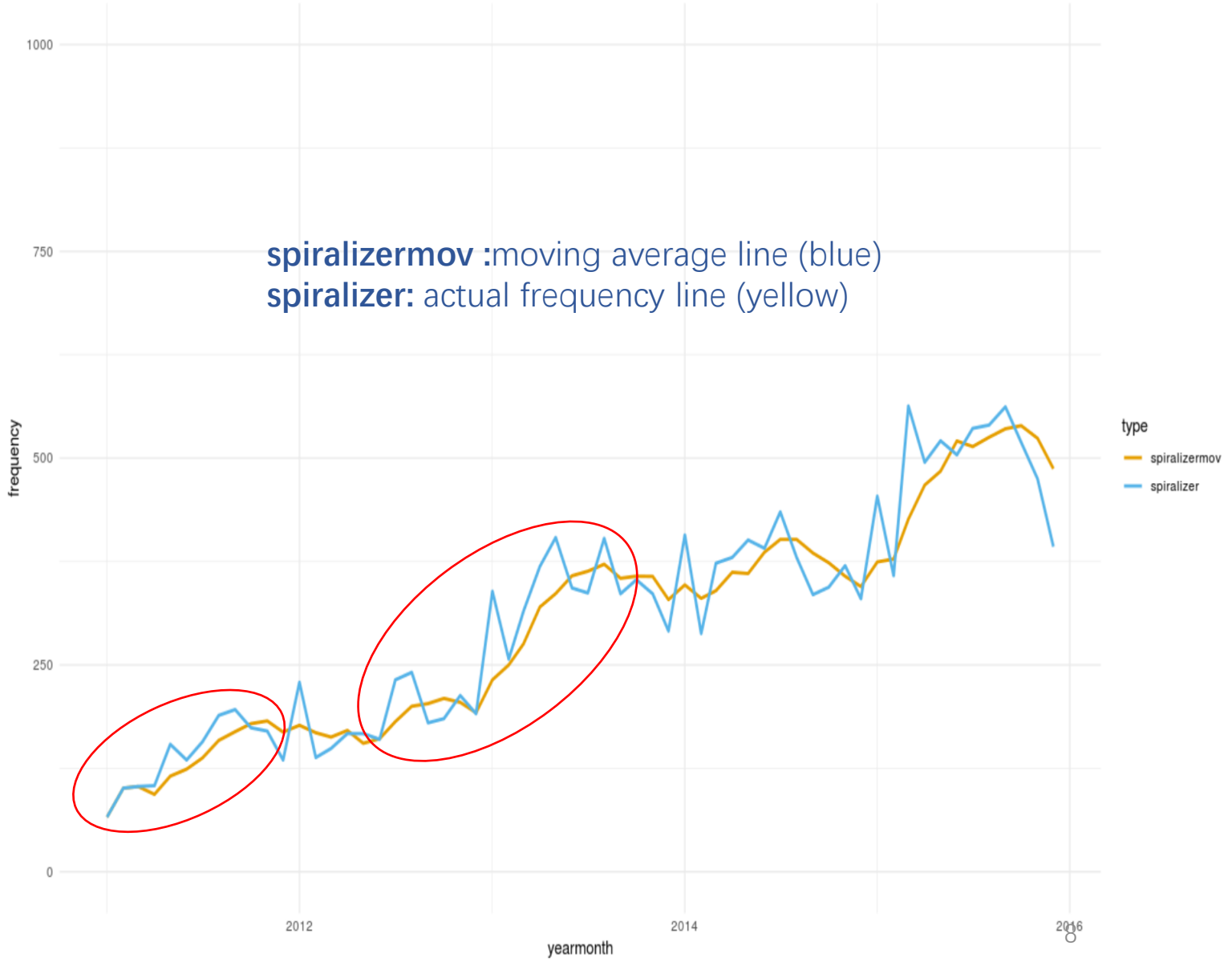
Analysis Demonstration for Vegetable Noodle Topic

Part 3. *Moving Average* frequency trend vs. Actual trend of the word "cauliflower".

- From the **middle of 2011**, the moving average frequency line of "spiralizer" has started to show a trend to be above the actual frequency line.
- From the **middle of 2012,** the trend has started to become more and more obvious, and there was several shot-term sharp increases of the topic around **2013~2014**, then the trend has been followed by public in **2015**, and has cooled down a little bit since the **late 2015**.

Speculation:

- Based on this chart, we can detect the trend as early as the **middle of 2011**, and we can validate it around **2013~2014.**



Trend Detection Methodology Wrap-up

Step 1. Extract terms to be explored.

- Perform topic modeling within food & ingredients vocabulary to detect a set of new or interesting topics.
- Extract the most representative terms within each selected topic via singular value decomposition (SVD) and co-occurrence analysis.
- *I didn't perform topic modeling in this assignment, but the following steps of analysis has been validated.*

Step 2. Explore frequencies of selected terms in time-series sequence.

- Calculate frequency and frequency proportion of selected terms based on DTM.
- Visualize the frequency and frequency proportion trends across time.
- Detect the most representative term with the least false positives based on frequency proportion trend.
- Analyze the trend from the chart and check whether there is any general rising pattern of a certain topic.

Step 3. Further explore the hidden pattern of a topic with moving average model.

- Select a fixed subset size of frequency data of the most representative term (selected in Step 2) to calculate the moving average of the original data.
- Visualize the moving average trend vs. the actual frequency trend of the representative term.
- Analyze the chart to detect the early emerging trend of a corresponding new topic.
- Track the data of the certain topic for a time span and estimate the continuous moving average to validate the trend.