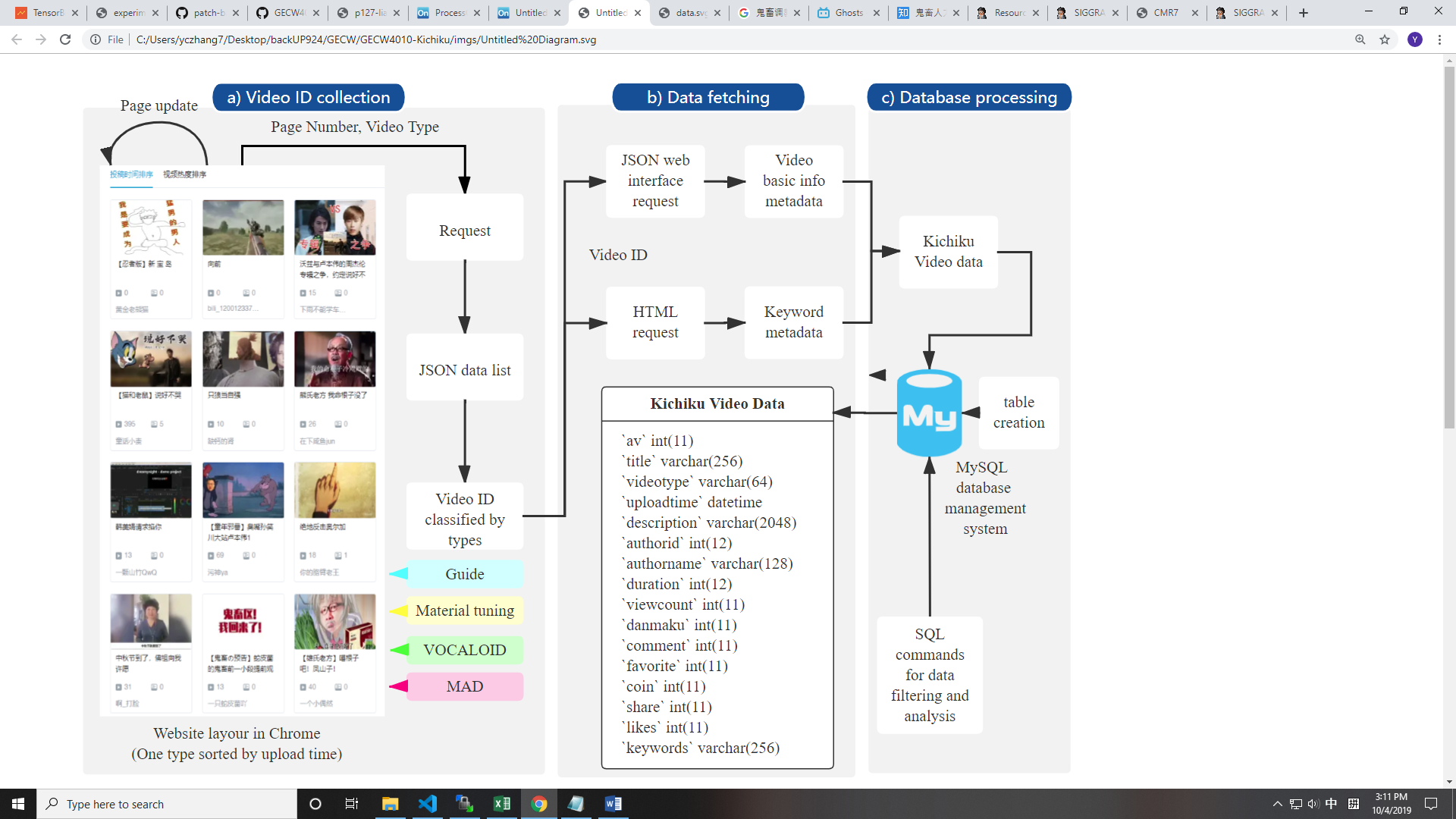
**Data Collection**

Adequate evidence is necessary to analyze culture around Kichiku and do further populairty measure and analysis. Thus, video data collection is the fundamental step to support our discussions. One of the most direct method to get correlated video data is fetching video data records from the website of online video platforms. This section mainly discusses about data fetching technique we applied to get Kichiku video meta-data from Bilibili, which has developed as the largest youth video-sharing platform, as well as one of the origin platform of Kichiku in mainland China in the last decade.

In this collection step, aiming on further analyze, besides of basic video ID (state as *AV-number* in Bilibili), we focus on three functionalities to decide the table format in MySQL database and items of collected data.

1. **Popularity measure metadata**, including count of *view, comment, danmaku, favorite, coin, share, like and duration*. There are multiple norms to define the popularity of a video. To setup a measure standard specialized on Kichiku video (details will be discussed in the *Popularity Analysis* part), these statistical data should be taken in our consideration.
2. **Producer-related metadata**, including *auhorid, authorname, fan count[[1]](#footnote-1)*. We find that the group of video producer has a significantly impact on popularity of Kichiku videos. These data records are used in our survey design and popularity analysis.
3. **Content-related metadata,** including *title, description, keywords, video subtype and uploadtime*. Relationship between video content and pupularity is the core part of this subculture study. Hot themes (tags from keywords) of a video change by time is a epitome of Kichiku changes in the past ten years.

We use Web data extraction technique to fetch data items mentioned above. The overall pipeline of the data collection process is shown in the fiure below. Comparing with traditional method, this process could be divided into 3 substeps in general. Firstly, directed search on each Kichiku subtype is applied to get video ID and save them into array-like files, this step is completed by sending request API which will return the newlist of a video displayed on the website. Secondly, the video information is fetched by requests with JSON and HTML replies, in which JSON response returns the keyword data. Thirdly, arranged data items are sent to MySQL database management system, stored there for invalid data filtering and analysis.



With our approach, 165380 valid items were returned as video results. From the timespan in September 2009 to May 2019. The cumulative data was captured in July 3rd, 2019. Two-month gap was set to avoid imperfection of data for popularity analysis with is accumulated with time such as view count and share count.

1. The fan count data is fetched in a separate step, which will be discussed in the data analysis part. [↑](#footnote-ref-1)