# Motivation[[1]](#footnote-1)

**Datasheets for Datasets  
*Adapted from:* *Gebru, Morgenstern, Vecchione, Vaughan,   
Wallach, Daumeé, and Crawford. (2018). Datasheets for Datasets.***

**Derive**

***1.1*** *For what purpose was the dataset created? Was there a specific task in mind? Was there a specific gap that needed to be filled? Please provide a description.*

Research conducted in the past suggests that YouTube tutorials have a positive impact on the skill development of younger people (Iftikhar, 2019). It should be noted that the creative segment like creating music was not covered here. To make this research possible, researchers will need a dataset. And that is where this dataset comes in to play focusing on the music industry. This dataset could be seen as a first step to perform studies on video attributes within the creative segment of YouTube. For what we could find on the internet, this data is not yet available.

For our dataset, we opted to focus on video tutorials about the program FLstudio. A program used to produce music mainly in the electronic music industry. Big names such as Martin Garrix started learning through this program and these artists are still using this to produce music that is on top of the music charts.

This program is widely known because of its simplicity which opened up opportunities for skill sharing through platforms like YouTube. The choice for YouTube was based mainly on the popularity of the program. The other big video program Tiktok was not selected because the videos are very short and can’t be defined as tutorials to get started as music producer. Vimeo was not selected because of popularity reasons.

The last choice to make was about using the API or to scrape the website. As with all websites it is possible to scrape YouTube, however YouTube also provides a pretty good API. For our dataset we opted for the API, because based on the API documentation this would be sufficient for our dataset. With the API we can get almost all the video information needed and we deemed the retrieval limit of 10.000 quotas a day sufficient for our project.

***1.2*** *Who created this dataset (e.g., which team, research group) and on behalf of which entity (e.g., company, institution, organization)?*

This dataset was created by Quinten de Putter, Jeroen Maagdenberg, Sam van de Ven and Tayfun Ozcan. They formed group 15 during the Online Data Collection Management course at Tilburg University.

***1.3*** *Who funded the creation of the dataset? If there is an associated grant, please provide the name of the grantor and the grant name and number.*

The YouTube API used for this dataset is free of charge. And we deemed the retrieval limit of 10.000 quotas sufficient per day. Because of that conducting the dataset was free. There was no associated grant.

# Composition

**Datasheets for Datasets  
*Adapted from:* *Gebru, Morgenstern, Vecchione, Vaughan,   
Wallach, Daumeé, and Crawford. (2018). Datasheets for Datasets.\****

**Derive**

***2.1*** *What do the instances that comprise the dataset represent (e.g., documents, photos, people, countries)? Are there multiple types of in- stances (e.g., movies, users, and ratings; people and interactions between them; nodes and edges)? Please provide a description.*

The datasets contain information about YouTube videos on the topic FLstudio tutorials. For each video the statistics are available in the dataset. This contains the video id, the publishing date, the channel id, video title, video description, channel title, video tags, the language of the video, the category id and if it was a live broadcast. We also extracted data per video, which contains the following variables: video-id, number of views, amount of likes and dislikes and the number of comments. We also extracted channel data, which contains the following variables; channel id, view count, subscriber count and video count. There are no interactions between the data variables.

***2.2*** *How many instances are there in total (of each type, if appropriate)?*

In the merged video dataset are 582 videos in total and we have a channel dataset that contains 264 channels.

***2.3*** *Does the dataset contain all possible instances or is it a sample (not necessarily random) of instances from a larger set? If the dataset is a sample, then what is the larger set? Is the sample representative of the larger set (e.g., geographic coverage)? If so, please describe how this representativeness was validated/verified. If it is not representative of the larger set, please describe why not (e.g., to cover a more diverse range of instances, because instances were withheld or unavailable).*

The dataset is limited and does not contain all possible instances since the API is bound to a certain quota limit of 10000 units. The larger set would be all possible instances that are on YouTube. With more time it might be possible to closer to the total number of videos that are tutorials towards FLstudio. We only managed to get this amount of data due to the limited timespan. Further research might want to increase this by letting the API run a longer time or use more computers to get more data.

As far as representativeness goes, the search results are not discrimated on language or geographical location. However, the number of observations could have been bigger. We recommend to get more data if someone wants to do research in the tutorial field. Our dataset could be used as a start or as a sample, but more data observations are needed for proper research. Especially since the only restriction for more data is time.

Someone can decide to pay to be able to get a higher extraction limit, more information about that can be found in the API documentation.

***2.4*** *What data does each instance consist of? “Raw” data (e.g., unprocessed text or images) or features? In either case, please provide a description.*

The instances that are gathered for the research consist of raw data. This means that the dataset contains unprocessed text and images. As mentioned in 2.1, information and statistics about each instance, or here video, are retrieved. Examples of the information and statistics are view count, like-count, channel name, channel id, thumbnail images, descriptions and similar results that one can find on a YouTube page directly but placed in a dataset. We also have channel data in another dataset, which contains all the data about the YouTube Channel. Total view count, amount of subscribers and video count are included in that one.

***2.5*** *Is there a label or target associated with each instance? If so, please provide a description.*

Each instance is uploaded by a YouTube channel. The target of the instance is to gain interaction in the form of views, likes etc.

The data can be linked together on two things: Channel ID and video id. By using those two labels someone can combine the three datasets together. One thing they might want to analyse if the dislike to like ratio differs over channels. Or that channels with more subscribers have a better like to dislike ratio.

***2.6*** *Is any information missing from individual instances? If so, please provide a description, explaining why this information is missing (e.g., because it was unavailable). This does not include intentionally removed information, but might include, e.g., redacted text.*

All the information of the videos that could be provided by the API is available in the dataset.

***2.7*** *Are relationships between individual instances made explicit (e.g., users’ movie ratings, social network links)? If so, please describe how these relationships are made explicit.*

Some things in the data can be seen as a relationship between variables. Such as views of a video and the total amount of views on a channel. If a video of a certain channel gets more views, the total number of views per channel also rises. The relation between views and comments is also pretty straightforward. When a video has more views, normally the number of comments also rises. Those simple, pretty straightforward relations are the only ones in this dataset.

***2.8*** *Are there recommended data splits (e.g., training, development/validation, testing)? If so, please provide a description of these splits, explaining the rationale behind them.*

We already chose to deliver multiple files instead of creating one big dataset to increase the readability of the dataset. It is however possible to link all three of them if the researcher wants to or first adjust the datasets and then combine them together.

So, there are no recommended splits since we already did that.

***2.9*** *Is the dataset self-contained, or does it link to or otherwise rely on external resources (e.g., websites, tweets, other datasets)? If it links to or relies on external resources, a) are there guarantees that they will exist, and remain constant, over time; b) are there official archival versions of the complete dataset (i.e., including the external resources as they existed at the time the dataset was created); c) are there any restrictions (e.g., licenses, fees) associated with any of the external resources that might apply to a future user? Please provide descriptions of all external resources and any restrictions associated with them, as well as links or other access points, as appropriate.*

The problem of extracting data through a video program like YouTube is that it can change day by day. Everyday there are videos published and the amount of likes, dislikes and comments can also change. The data in this dataset is just a snapshot. The data is also depending on the YouTube algorithm, so the guarantee that it will remain constant can’t be guaranteed. As far as official archival versions go, as far as we know there aren’t any of those. The only extraction restriction is the extraction limit of 10000 units a day. Future users will also have this problem. However, this quota limit can be increased by buying certain packages/subscriptions on the Google Cloud Platform. So for this questions we can’t give any external resources or description as there aren’t any for our dataset, it is self-contained.

***2.10*** *Does the dataset contain data that might be considered confidential (e.g., data that is protected by legal privilege or by doctorpatient confidentiality, data that includes the content of individuals non-public communications)? If so, please provide a description.*

The raw data collected for this research is publicly available. Therefore, the dataset is not considered confidential. People that upload a video can choose their own channel name and thus a large number of channel names in our dataset are nicknames or company names.

***2.11*** *Does the dataset contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety? If so, please describe why.*

The dataset contains only raw data about YouTube videos and provides information and statistics about these videos. These statistics are not offensive, insulting, threatening and will not cause anxiety since each video has to comply to the YouTube guidelines and policies.

***2.12*** *Does the dataset relate to people? If not, you may skip the remaining questions in this section.*

The dataset does not relate to people. Therefore, the questions from 2.12 till 2.15 will not be answered.

***2.13*** *Does the dataset identify any subpopulations (e.g., by age, gender)? If so, please describe how these subpopulations are identified and provide a description of their respective distributions within the dataset.*

N/A

***2.14*** *Is it possible to identify individuals (i.e., one or more natural persons), either directly or indirectly (i.e., in combination with other data) from the dataset? If so, please describe how.*

N/A

***2.15*** *Does the dataset contain data that might be considered sensitive in any way (e.g., data that reveals racial or ethnic origins, sexual orientations, religious beliefs, political opinions or union memberships, or locations; financial or health data; biometric or genetic data; forms of government identification, such as social security numbers; criminal history)? If so, please provide a description.*

N/A

# 3. Collection Process

***3.1*** *How was the data associated with each instance acquired? Was the data directly observable (e.g., raw text, movie ratings), reported by subjects (e.g., survey responses), or indirectly inferred/derived from other data (e.g., part-of-speech tags, model-based guesses for age or language)? If data was reported by subjects or indirectly inferred/derived from other data, was the data validated/verified? If so, please describe how.*

The collected data consists of raw descriptive data and statistics for each instance. A large part of the data was not directly observable, since it was raw text. However, some information such as titles, descriptions, view count, like- and dislike-count are directly observable. The data was not reported by subjects and not derived from other data sources.

***3.2*** *What mechanisms or procedures were used to collect the data (e.g., hardware apparatus or sensor, manual human curation, software pro- gram, software API)? How were these mechanisms or procedures validated?*

As mentioned before, for this research, a YouTube API was used to gather data. This API was developed by Google itself and, for access, it is required to use an API key that is requested through the Google Cloud Platform.

***3.3*** *If the dataset is a sample from a larger set, what was the sampling strategy (e.g., deterministic, probabilistic with specific sampling probabilities)?*

YouTube allows researchers a maximum of 10.000 quotas to gather raw video data. Since there is an endless number of videos about FL studio tutorials, it obligates the researchers to work with a sample that fits the quota limit. Also, since the purpose of this research is to find out which FL studio tutorial videos are the most popular, the data consist of the most relevant videos.

***3.4*** *Who was involved in the data collection process (e.g., students, crowdworkers, contractors) and how were they compensated (e.g., how much were crowdworkers paid)?*

The people who were involved are four students from Tilburg University (Quinten de Putter, Jeroen Maagdenberg, Sam van de Ven and Tayfun Ozcan). They were not financially compensated for the collection process. The process was initiated and regularly reviewed by professor, dr. Hannes Datta from Tilburg University.

***3.5*** *Over what timeframe was the data collected? Does this timeframe match the creation timeframe of the data associated with the instances (e.g., recent crawl of old news articles)? If not, please describe the time- frame in which the data associated with the instances was created.*

As mentioned in section 2.9, the data is a snapshot of the most relevant results for the search query “FL tutorial” at a given moment. Therefore, there is not a specific timeframe wherein the data is collected.

***3.6*** *Were any ethical review processes conducted (e.g., by an institutional review board)? If so, please provide a description of these review processes, including the outcomes, as well as a link or other access point to any supporting documentation.*

There were no ethical review processes conducted.

***3.7*** *Does the dataset relate to people? If not, you may skip the remaining questions in this section.*

N/A

***3.8*** *Did you collect the data from the individuals in question directly, or obtain it via third parties or other sources (e.g., websites)?*

N/A

***3.9*** *Were the individuals in question notified about the data collection? If so, please describe (or show with screenshots or other information) how notice was provided, and provide a link or other access point to, or other- wise reproduce, the exact language of the notification itself.*

N/A

***3.10*** *Did the individuals in question consent to the collection and use of their data? If so, please describe (or show with screenshots or other information) how consent was requested and provided, and provide a link or other access point to, or otherwise reproduce, the exact language to which the individuals consented.*

N/A

***3.11*** *If consent was obtained, were the consenting individuals provided with a mechanism to revoke their consent in the future or for certain uses? If so, please provide a description, as well as a link or other access point to the mechanism (if appropriate).*

N/A

***3.12*** *Has an analysis of the potential impact of the dataset and its use on data subjects (e.g., a data protection impact analysis) been conducted? If so, please provide a description of this analysis, including the outcomes, as well as a link or other access point to any supporting documentation.*

N/A

# 4. Preprocessing, cleaning, labeling

***4.1*** *Was any preprocessing/cleaning/labeling of the data done (e.g., discretization or bucketing, tokenization, part-of-speech tagging, SIFT feature extraction, removal of instances, processing of missing values)? If so, please provide a description. If not, you may skip the remain- der of the questions in this section.*

Within this project, there is no preprocessing, cleaning or labeling done. Therefore, the remainder of the section will be skipped.

On a side note, YouTube adjusts the date and time of a video upload to the local time of a person that accesses a video on the website or the API.

***4.2*** *Was the “raw” data saved in addition to the preprocessed/cleaned/labeled data (e.g., to support unanticipated future uses)? If so, please provide a link or other access point to the “raw” data.*

N/A

***4.3*** *Is the software used to preprocess/clean/label the instances available? If so, please provide a link or other access point.*

N/A

# 5. Uses

***5.1*** *Has the dataset been used for any tasks already? If so, please provide a description.*

No, for this research only data has been collected. Therefore, no tasks were performed.

***5.2*** *Is there a repository that links to any or all papers or systems that use the dataset? If so, please provide a link or other access point.*

There is no repository for this research.

***5.3*** *What (other) tasks could the dataset be used for?*

With our conducted dataset researchers could do simple things like multiple linear regression with different independent and dependent variables for their research, however this is pretty standard. Perhaps if researchers have a way to conduct data about the comment section and link that to our dataset they can do more complicated test analysis to check if the tutorial videos for example really help to improve other people their skill. They might also use the dataset for financial videos on YouTube. However, that research is not really limited to Flstudio tutorials.

With the YouTube data researchers can up with more advanced models to predict if a video is going to get a lot of views or not. Which is also more general.

If a researcher can get more data extracted from the YouTube API instead of our limited number of extractions analysts can estimate how big the request from customers (viewers) is in this segment, which is good to know for advertisers and Google, so they can improve their sales and revenue. An addition on that might be to check how some YouTube Channels perform. What also can be done is to see if the like to dislike ratio decreases or increases in more popular videos compared to videos with less videos.

There are multiple things a researcher can do with this data. What also can be seen as a valuable research is that YouTube can become a good place to gather data from and use that for research purposes. One can use our API code and use that to extract data from other topics by adjusting just a couple things in the script.

***5.4*** *Is there anything about the composition of the dataset or the way it was collected and preprocessed/cleaned/labeled that might impact future uses? For example, is there anything that a future user might need to know to avoid uses that could result in unfair treatment of individuals or groups (e.g., stereotyping, quality of service issues) or other undesirable harms (e.g., financial harms, legal risks) If so, please provide a description. Is there anything a future user could do to mitigate these undesirable harms?*

It should be noted that the data is just a snapshot. Videos get uploaded every single day and the number of views and likes can also change every moment. This is a serious limitation for research. Secondly, by extracting data this way we are depending on the logarithm which can differ per person if someone is extracting data.

A future user of our data should take into account the three datasets and before continuing select the variables he or she needs and combine the data together. If the user wants more data, he probably should upgrade to the premium service so he or she can extract the data faster.

***5.5*** *Are there tasks for which the dataset should not be used? If so, please provide a description.*

The data might contain real and fake first and last names. Therefore, it is not advised to use the names gathered as a reference for contacting certain people.

1. \* https://arxiv.org/abs/1803.09010 [↑](#footnote-ref-1)